Data Mining and Machine Learning to Predict Acute Coronary Syndrome Mortality

by

Juliana Binti Jaafar

Submitted in accordance with the requirements for the degree of Doctoral Degree (Ph.D)

The University of Leeds
Leeds Institute of Health Sciences

September, 2017

Declarations

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some parts of the work presented in this thesis have been published in the following article:

JAAFAR, J., ATWELL, E., JOHNSON, O., CLAMP, S. & AHMAD, W. A. W. 2013. Evaluation of Machine Learning Techniques in Predicting Acute Coronary Syndrome Outcome. *Research and Development in Intelligent Systems XXX.* Springer.

The publication contains the preliminary studies discussed in Chapter four. The candidate confirms that the above jointly authored publications are primarily the work of the first author. The role of the second author was editorial and supervisory.

Acknowledgements

Alhamdulilah!

My most sincere and greatest appreciation goes to my supervisors,

***Professor Jeremy C Wyatt, Dr Eric Atwell, Dr Susan Clamp, and Owen Johnson**,*

for their immense support, valuable advice, and unfailing guidance along the journey. This research would not have been possible without them.

My gratitude extends to:

***The Malaysian National Cardiovascular Disease Database (NCVD)**team, and especially Professor Dr. Wan Azman Wan Ahmad. Thank you for your endless support and advise pertaining to the Malaysian datasets. I am truly looking forward to more collaboration on the NCVD dataset for the benefit of the Malaysian health care industry as a whole, and, specifically, overall cardiac services*

***The Improving Prevention of Vascular Events in Primary Care (IMPROVE-PC)** project team and Steve Magare, former data manager of IMPROVE-PC. Thank you for granting access to the dataset and for your continual guidance to better help mecomprehend the dataset.*

***The University of Kuala Lumpur (UniKL)**, for the sponsorship.*

***Leeds Institute of Health Science - The Yorkshire Centre for Health Informatics (YCHI)** , and to **the School of Computing**, for giving me the opportunity and facilitating my continuous learning process.*

*To my **dearest friends**, all my **great colleagues,** and the **amazing group of people** who studied alongside me throughout my years at the University of Leeds - thank you for keeping me sane throughout my PhD journey!*

*Finally, my greatest-beyond words, 'thank you,' to my **parents, husband, sister, and brothers,** for their continuous prayers, love, encouragement, and moral support. I would have not been able to do this without all of youby my side! And to my **daughter**, your existence in this world has made me stronger than ever!*

Abstract

This thesis has investigated and demonstrated the potential for developing prediction models using Machine Learning(ML) algorithms on registry datasets. Many current Acute Coronary Syndrome (ACS) prediction models, were developed using traditional statistical methods. In an era of big-data evolution, ML offers a spectrum of algorithms that aid in generating prediction models for ACS. This study has explored 29 algorithms with which to build ACS prediction models for Asian (Malaysia) and Western (Leeds, UK) registries, covering patients withal types of ACS and those with the new standard ACS treatments. The internal and external validation of the models present satisfactory calibration measures, indicating the ability of ML algorithms to produce competitive models in comparison to traditional statistical methods.

To achieve simpler, yet competitive predictive performance, comprehensive ML feature selection methods have been evaluated, and Correlation-Based-Feature-Selection(CFS) emerged as the best method. This thesis also has evaluated the potential of predictors of existing ACS models to be adapted to other registries' data. Despite different regions and different population characteristics, most of the existing predictors remains constant with the outcome. Thus, the findings suggest that, with some adjustments customized to the registry, the existing predictors can be adopted to develop a simple model and expedite the model development process. Furthermore, the strength of the predictors of each clinical categories has also been evaluated. The results suggest that, to construct a satisfactory ACS model, combination of predictors from various clinical events is essential. At the very least, to achieve a satisfactory model, combination of demographic, medical history, and clinical presentation information categories is required. However, predictors from medication history category has found to be worthless in terms of contributing to a better prediction model.

Next, this study has investigated classifier degradation in ML model development. The findings suggest that the overlapping instances in minority class of imbalanced dataset and missing values are the main problems of classifier degradation.

New methods i.e. the *overlapped-undersampling* method to handle imbalanced dataset and the *mean-clustering-imputation* method to handle missing values have been introduced. The *overlapped-undersampling* failed to boost the model performance of the datasets. Nevertheless, the results suggest that more training samples on imbalanced datasets are sufficient to produce satisfactory models. The *mean-clustering-imputation* method produced better models compare to the simple imputation method and imputation method embedded in an algorithm. However, removing instances with missing data resulted in superior models.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACE	Angiotensin converting enzyme
ACS	Acute Coronary Syndrome
ADT	Alternating Decision Tree
AMIS	Acute Myocardial Infarction in Switzerland
ANN	Artificial neural networks
BB	Beta Blocker
BBB	Bundle branch block of ECG
BFT	Bloom-Filter Tree
BP	Blood pressure
BS	Brier score
BN	Bayes Net
CABG	Coronary artery bypass grafting
CFS	Correlation-Based-Feature Selection method
CHD	Coronary Heart Disease
CR	Conjunctive Rule
CVD	Cardiovascular disease
C-ACS	The Canada Acute Coronary Syndrome Risk Score
DBP	Diastolic Blood Pressure
DM	Data Mining
DM-ML	Data mining using machine learning
DS	Decision Stump
DSS	Decision Support System
DT	Decision Tree
DTNB	Decision Tree and Naive Bayes
ECG	Electrocardiogram
EHR	Electronic health record
EMMACE	Evaluation of Methods and Management of Acute Coronary Events
FBG	Fasting blood glucose
FT	Functional Tree
GP	General Practices
GRACE	Global Registry of Acute Coronary Events

GUSTO	The Global Use of Strategies to Open Occluded Coronary Arteries
HES	Hospital Episode Statistics
IMPROVE-PC	Improving Prevention of Vascular Events in Primary Care
Jrip	Repeated Incremental Pruning to Produce Error Reduction
J48	C4.5 Decision Tree
J48Graft	Grafted C4.5 Decision Tree
KNN	K-Nearest Neighbour
LBBB	Left bundle branch block of ECG
LDL-C	Low-density lipoprotein cholesterol
LG	Logistic Regression
LMWH	Low molecular weight heparin
LT	Logistic Alternating Decision Tree
LMT	Logistic Model Tree
Lvef	Left ventricular ejection fraction
LWL	Locally Weighted Naive Bayes
MACE	Major Adverse Cardiovascular Event
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MI	Myocardial infarction
MINAP	Myocardial Ischemia National Audit Project
ML	Machine Learning
MLP	Multi Layer Perceptron
MR	Multinational registry
NB	Naive Bayes
NBT	Naive Bayes Tree
NCVD	National Cardiovascular Disease Database
NSTEMI	Non-ST-elevation myocardial infarction
PCI	Percutaneous Coronary Intervention
RCT	Randomized control trial
REPT	Reduces Error Pruning Tree
RF	Random Forest
RT	Random Tree
RR	Regional registry

OneR	One Rule
PART	Projective Adaptive Resonance Theory
Ridor	RIpple-DOwn Rule
SBP	Systolic Blood Pressure
SC	Simple Cart
SD	Vanderbilt University Medical Centre's Synthetic Derivative
SMOTE	Synthetic Minority Over-Sampling Technique
SRI	Simple Risk Index
STEMI	ST-Elevation Myocardial Infarction
SVM	Support Vector Machine
TIMI	Thrombolysis in Myocardial Infarction
UA	Unstable angina
WEKA	Waikato Environment for Knowledge Analysis
ZR	ZeroRule

Chapter 1: Introduction

This chapter presents the research background, objectives, and thesis overview. It begins with an overview of Acute Coronary Syndrome (ACS), the burden of ACS worldwide, and the use of risk prediction models for ACS. Next, the chapter discusses the research problem and motivation for the study. It ends with an elaboration on the thesis structure.

1.1. Acute Coronary Syndrome (ACS)

Acute Coronary Syndrome (ACS) refers to a heart condition that is caused by the blockage of the blood supply to the heart, and is commonly due to the development of plaque inside the arteries (*atherosclerosis)*. Blockages in the arteries reduce the oxygen supply to the heart, resulting in the sudden onset of angina (unstable angina (UA)) to severe chest pain, and subsequently, damaging the heart (myocardial infarction (MI) or acute MI). This life-threatening condition can become more severe if no early invasive management strategy is performed to restore blood flow to the heart. .

Chest pain or pressure in the chest is a vital symptom of ACS. Other symptoms include sweating; dizziness or fainting; difficulty in breathing; pain or feeling pressure or a strange feeling in the back, neck, jaw, or either arm; and fast or irregular heartbeat. In addition, several existing risk factors also increase the possibility of ACS among patients. Moreover, the risk factors of ACS are mainly the risk factors of cardiovascular disease (CVD), which include both non-modifiable risk factors, such as older age, being male, family history of CVD, ethnicity, and modifiable risk factors. Some well-known modifiable risk factors include excessive total cholesterol, obesity, diabetics, hypertension and stress, smoking status, and a low-quality lifestyle(2009, Philip I. Aaronson and Ward., 2007., Swales and P. De Bono, 1993).

The diagnosis of ACS in a patient begins with a thorough assessment of visible symptoms, an electrocardiogram (ECG), and measurements of cardiac biomarker levels. In addition, a patient's past medical history and

existing risk factors are essential in helping to both diagnose and manage ACS.

The varied categories of ACS, i.e., ST-elevation myocardial infarction (STEMI), non-STEMI(NSTEMI), and UA, distinguish the treatment and intensity of therapeutic intervention. In fact, patients diagnosed with STEMI need immediate intervention to restore blood supply to the heart, as such a severe condition could lead to death. After the condition has been stabilized, the next step is to prevent recurrence. Percutaneous coronary intervention (PCI) and fibrinolytic therapy are the standard treatments for patients with STEMI(Smith et al., 2015), while other coronary reperfusion may include coronary artery bypass graft (CABG). On the other hand, patients with NSTEMI and UA are treated to stabilize and limit the progression of ischemic events. Moreover, early invasive treatment is necessary for patients with a higher risk of NSTEMI/UA.

1.2. Global Burden of Acute Coronary Syndrome (ACS)

ACS is a subgroup of coronary heart disease (CHD). CHD is a leading cause of death worldwide, with 7.3 million deaths recorded in 2001(Gaziano et al., 2010). Nonetheless, the mortality rate for CHD has witnessed a decreasing trend in North America and many Western countries due to better prevention, diagnosis, and treatment, as well as changes made towards a healthier lifestyle. Despite the decrease in mortality rate, it is still a major cause of morbidity, and a single cause of death in many nations. One out of five deaths in the United States of America(USA) were caused by CHD in 2005(Lloyd-Jones et al., 2009). In addition, CHD also contributes to premature death in most European countries. For instance, 20% of males and 18% of females below the age of 75 died of CHD in 2002 (Allender et al., 2008). On top of that, there is evidence that depicts a growing burden of CHD in developing nations, mainly due to rapid economic development and social transformation (Gaziano et al., 2010).

Furthermore, the high rates of morbidity and mortality have become a major economic burden factor. More practice guidelines are required to provide treatment, care, and support options for managing overall CVD

treatment. Apart from the increasing cost in health care, such as treatment, medication, and prevention, countries with a high CHD rate also have to bear the loss of economically productive resources due to the inability to work and premature deaths. For example, In 1995, Germany spent USD 26 billion for the direct cost of hospital care and rehabilitation, whereas it lost USD 48 billion due to lost productivity caused by short- and long-term disabilities, as well as short-term deaths attributed to CHD(Mackay et al., 2004). Meanwhile, Abegunde et al.(2007)discovered that the burden of and economic loss due to stroke, heart disease, and diabetes among 23 low- and middle-income countries was estimated to be USD 84 billion dollars between 2006 and 2015 if no effort was taken to reduce the risk of overall CVD.

Such concerns have led to constant efforts to improve all spectra of ACS treatment, care, and prevention. The drastic improvement of CHD and overall CVD death rates in some regions has been due to prevention and treatment (Roth et al., 2015). Better CHD management, early identification of high-risk patients, changes in lifestyle, and public awareness are some cost-effective strategies for managing the burden.

1.3. Prediction Models and Uses of Risk Prediction Models in ACS

Risk prediction models, clinical prediction models, prediction and prognostic models, prediction rules, and risk scores refer to a range of terminologies used to describe a model used to predict an outcome. In this particular thesis, the notion ‘prediction model’ is utilized. A prediction model employs a set of predictors in predicting the presence (diagnosis) or the occurrence of a certain outcome (Toll et al., 2008). For reference purposes, risk factors generally describe the factors that are causally related to getting the disease. Meanwhile, predictors in a prediction model refer to the attributes found in a prediction model that may causally relate to the outcome, but not necessarily. For example, cigarette smoking is associated with an increased risk of cancer (good predictor), but has never been actually proven to cause cancer.

Prediction models are generally constructed by using standard statistical regression approaches, i.e., multiple linear, multiple logistic, or Cox regression. The regression approach, in general, has been employed in clinical prediction for some time because it generates simple correlations between the predictors and outcome. In particular, logistic regression (LG) has become a familiar approach that can be easily implemented in most statistical packages, such as Statistical Packages for Social Science (SPSS), STATA, and R. Linear regression is normally used when dealing with a continuous outcome, whereas LG is commonly used to predict a binary outcome. On the other hand, Cox or proportional hazard regression is normally applied when the effect of the variables is evaluated for a certain period of time. Hence, both linear and LG approaches are usually used for diagnostic or short-term prognostic model, e.g., predicting in-hospital mortality, while Cox regression is widely used for predicting long-term events, e.g. predicting a 10-year mortality rate.

Techniques based on Machine Learning(ML) have also been used in constructing prediction models. Unlike statistical modelling, ML algorithms are used to learn the datasets for prediction purposes. Data Mining (DM) using ML (DM-ML) provides a spectrum of learning algorithms comprised of non-linear methods, linear methods, and ensemble methods (a combination of multiple learning techniques), which may better describe the relationships between the identified predictors in building prediction models. In comparison to the classical statistical method, although the ML approach does not appear to be a popular choice for clinical prediction modelling, the potential of these ML algorithms has been established in developing prediction models. Two of the most popular ML techniques that have been explored, particularly in ACS prediction models, are the artificial neural networks (ANN) (Baxt et al., 2002;Harrison and Kennedy, 2005;Green et al., 2006b;McCullough et al., 2007;Bassan et al., 2004) and decision-tree-based algorithms (Karaolis et al., 2010, Lavesson et al., 2009, Fonarow et al., 2005).

Randomized control trials (RCTs) and registries are where the data is to be found when constructing prediction models. The well-known ACS prediction models are generally derived from randomized controlled data

(Antman et al., 2000, Boersma et al., 2000, Lee et al., 1995, Morrow et al., 2001). In RCT study design, more accurate and complete samples are collected within a defined prospective, method, and duration. Nonetheless, these samples may not represent the real population or scenario, thus affecting the generalisability of the model derived from clinical trials. Moreover, bias may exist, as the samples for RCT are selected based on certain defined inclusion and exclusion criteria. In real hospitals or healthcare centres, the practice methods and duration are varied and may not strictly adhere to a defined standard procedure.

The advent of the electronic health record (EHR) within healthcare organizations has shifted the methodology of deriving, building, and validating prediction models(Cooney et al., 2009;Granger et al., 2003;Hippisley-Cox et al., 2008). At present, the EHR-based registries are an essential asset in building prediction models. An EHR is a collection of an individual patient's records which comprehensively records both clinical information and administrative information in digital format. Example of information collected by EHRs are medical histories, progress notes, medications/prescriptions, laboratory data and radiology reports, billings, and appointments. On the other hand, a registry is a collection of information collected with defined purpose(s) and populations to observe a specified health outcome(Gliklich et al., 2014). Many registries are derived from EHRs, hence imposing the challenges of working with EHRs data. The main disadvantage of working with EHR-based registry data is the data quality issue. More missing data, noise, and dirty data can be expected when working with this registry data. Thus, a longer time is needed to prepare the data for modelling. However, with proper strategies in data cleaning and transformation, this registry data can be advantageous in constructing a reliable prediction model.

The ACS prediction model aids clinicians in identifying patients at high risk for mortality following ACS events. Patients with high risk can properly be advised with adequate and on-time treatment, whilst patients with low risk can be assigned less aggressive treatment. In addition, a reliable prediction model and categorization of summary measures could aid in the realization

of stratified medicine to provide individualized treatment and interventions according to individual needs.

A prediction model can also suggest better resource management, result in cost-savings, and minimise unnecessary treatment complications (Lloyd-Jones, 2010, SIGN, 2007). Moreover, utilizing prediction model to identify patients with high risk has become an established practice for a cost-effective and primary prevention strategy (Bassand et al., 2007, SIGN, 2007 (Updated 2013), Gaziano et al., 2010). Furthermore, prediction models that estimate long-term outcomes are useful in managing the long-term care of high-risk patients.

Prediction models also provide prognostic information that is valuable not only to clinicians, but also for patients and their families. Understanding the risk level, as well as the methods of prevention and care, can help in managing ACS events and preventing reoccurrence.

Furthermore, identification of patients with varying risk levels also helps in clinical trial analysis and epidemiological studies for it offers information for the examination of treatment effects by the varied risk levels of patients. On top of that, it is also helpful in regulating baseline risk factors, as well as screening for both inclusion and exclusion in clinical trials.

Hence, a prediction model provides a significant contribution to vital prevention strategies, in addition to formulating effective treatment in clinical practise and resource management. Thus, practical, reliable, and accurate prediction models are indeed helpful for medical decision-making in effectively implementing prevention strategies. Moreover, a reliable prediction model exploits suitable techniques and strategies to build a model, which can be quantified by using valid performance measures and derived from an appropriate number of samples and risk factors (Cooney et al., 2009, Lloyd-Jones, 2010).Nevertheless, a prediction model or any tool that is derived from the model can never substitute for the decisions of or judgements made by clinicians.

.

1.4. Background : Data Mining (DM) and Machine Learning (ML)

Data mining (DM) refers to the process of extracting useful knowledge from a large dataset by analysing not only the patterns, but also the correlations between its attributes (Han and Kamber, 2001). It also offers a platform for prediction modelling, in addition to revealing ‘hidden’ knowledge. It employs intelligent techniques and structural methods to not only unravel and describe the pattern, but also to evaluate the pattern in accordance to several measures.

Machine learning (ML), on the other hand, denotes a multidisciplinary field of artificial intelligence, statistics, probability, computational complexity theory, information theory, learning theory, and other numerous fields (Meyfroidt et al.,2009), which is focused on the development of algorithms/techniques to learn from experience in the attempt to improve the performance of a system over time (Mitchell, 1997).The aims of ML are to allow for automatic 'learning' of data using computer algorithms and make generalizations on what has been learnt to new, but similar data. Learning can be categorized into supervised and unsupervised learning. Supervised learning identifies an approximate mapping function of input data to the output variable. The tasks of supervised learning include classification/prediction, regression, and feature selection. In contrast, unsupervised learning focuses more on learning the data by understanding the patterns in the data. Tasks associated with unsupervised learning include clustering and making associations (Witten et al., 2005, Weiss and Indurkhya., 1998)

In the context of DM, ML offers intelligent techniques and methods for mining the data, i.e., for discovering and describing patterns while focusing on inductive learning (learning by example). ML can be considered to be the core of DM. In terms of predictive modelling, unlike traditional statistics, which relies on small samples with pre-defined assumptions on data and its distribution, ML models the data on a given task through an heuristic approach with minimal pre-assumptions about the data and problem

Advancements in software and hardware have further allowed vast

amount of data to be captured and stored, ranging from simple to complex data types. This scenario has created the need to intelligently process data, identify interesting patterns, and transform this into useful information and knowledge. Therefore, extracting the right and meaningful information is vital especially in acquiring economic advantage. As such, DM has appeared to be a major evolution in the information industry. Massive data have also encouraged the development and the evolution of DM applications. Moreover, DM has been applied in many areas, such as marketing and retail, finance and banking, engineering, sports, as well as the medical and health industry. For instance, DM has been employed to improve a present business process or the quality of a product in the attempt to anticipate future trends in planning strategies, predicting and identifying risks, formulating prevention measures, interpreting images, and recognising patterns(Kantardzic and Zurada, 2005, Han and Kamber, 2001, Wang et al., 2012, Choudhary et al., 2009, Delen et al., 2012a, Witten et al., 2005).

1.5. Background : DM in the Medical Field

The growth of medical data has encouraged the application of DM in the medical field. Outcomes from vast studies in medical DM have highlighted a great potential for improving efficiency and cost-saving aspects in clinical administration, as well as in clinical treatment and care (Koh and Tan, 2011). For example, the prediction model built by Zhong et al. (2012)applied a new hybrid DM using ML (DM-ML) technique, which exemplified the potential to improve the management of costs and budgeting for hospital administrations. Additionally, Chazard et al. (2011)and Bate et al.(2008)used DM to determine adverse drugs events.

Furthermore, DM possesses the ability to aid in making decisions for prognosis and diagnosis, as well as for treatment options(Pogorelc et al., 2012, Delen et al., 2005). DM-ML techniques have also been used in various developments of the Decision Support System (DSS) for rule generation (Tenorio et al., 2011, Del Fiol and Haug, 2009). In a typical medical DSS, rules are generated based on expert knowledge. However, via DM, rules are generated by the system and, later, validated by the domain expert, thus promoting efficiency in developing the system.

DM has also made a contribution to evidence-based medicine (Stolba and Tjoa, 2006). Evidence-based medicine is the concept of making decisions related to patients' health care that integrate clinical expertise, patients' values, and the best external evidence(Masic et al., 2008). The concept is an application of providing better health care and improving cost effectiveness. The knowledge extracted from a large and complex healthcare dataset turns out to be important evidence that should not be neglected in acquiring the best external evidence. For instance, Delen et al. (2010) applied DM to identifying more intricately predictive factors in estimating survival time after transplants. Other than that, Chu et al. (2009) utilized the Bayes model in presenting an evidence-based expert system to detect CAD from hospital-based data and existing epidemiological study. The evidence based expert system using the Bayes model has achieved 0.86 AUC rate on hospital based data and 0.86 on existing epidemiological study.

The capabilities of DM in handling huge datasets with tolerable performance time have encouraged the continuous employment of DM in medical studies(Delen, 2009, Sampson et al., 2011) . For example, Sitar-Taut et al.(2009)ranked the significance of identifying risk factors for coronary artery disease (CAD), stroke, and peripheral artery disease (PAD), which concluded that varied CVDs have varying ranks of important risk factors. Meanwhile, Khalilia et al.(2011)used the Mean Decrease Gini measure to determine the essential variables linked to various diseases, whereas Delen et al. (2012b)employed the sensitivity analysis method to identify the most important variables that had an impact upon outcomes of CABG surgeries.

In addition, DM-ML can also be combined with conventional statistical methods to produce more commendable estimations. As such, Tham et al.(2003) used ANN to predict CHD by combining a set of gene marker attributes with some typical risk factors as predictors. In order to identify the gene markers' input for ANN, several statistical methods, i.e., principal components analysis (PCA) and factor analysis (FA), were used.

Although a number of studies have demonstrated the potential of DM, such as by exhibiting potentially good predictive models or establishing new knowledge in the medical field, several works, such as work done by Sami

(2006), have been rejected by the medical community. Sami claimed that the work had been rejected due to the limited number of journals in the area of Urology that would consider the DM method that he used. Medical DM, thus, is considered unique because it inherits the complexity of medical data and its domain. Medical data is huge and heterogeneous. It is available in many forms, i.e., images, text, ECG readings, or even structured data, and comes from different events (databases/systems) such as administration, diagnosis, treatment, and even follow up. Moreover, there are a large number of missing values, inconsistencies, and imprecise and incomplete data in a medical dataset(Cios and Moore, 2002). A substantial number of studies have been initiated to investigate the uniqueness of medical DM. In fact, some examined issues surrounding medical DM, such as, privacy, ethics, and confidentiality issues; medical data intricacy and quality; and the exclusivity of medical approaches(Shillabeer and Roddick, 2007, Cios and Moore, 2002, Sami, 2006, Iavindrasana et al., 2009).

Thus, mining medical data means being able to handle the uniqueness of medical datasets. The medical field is exclusive in its approach as it deals with life or death, which applies to everybody. For example, DM, in general, concerns itself with digging out patterns and trends in a dataset. In contrast, in medical research, more concern is placed on the minority events, such as mortality, which does not conform to patterns and trends. As such, it is vital to conform to the medical paradigm in terms of measuring the error rate, i.e., the sensitivity and specificity measures of a prediction model, rather than measuring the accuracy rate as in general DM applications. Another example is reporting the DM results from a medical dataset. Cautious consideration of the language used is important as any information distortion in reporting medical results has the potential to be life threatening and have cost and political consequences. Appropriate and precise descriptions of the data source with detailed and defined characteristics of populations must be adhered to. This is indeed essential in a clinical research setting. Bouwmeester et al. (2012)concluded that many prediction models available in high-impact journals have very limited applicability due to not following the recommended methodologies.

1.6. Problem Statement

As mentioned previously, current prediction models are generally constructed by using standard statistical regression approaches, i.e., multiple linear, multiple logistic, or Cox regression. In this conventional way, the predictors of the models are generally identified from a set of candidate variables defined based on clinical expert opinions, risk factor findings from clinical studies, and/or previously developed models(Morrow et al., 2000, Boersma et al., 2000, Granger et al., 2003). Univariate LG and multiple LG are then run on these potential predictors to identify the significance predictors of the model. With ML, the paradigm of predictor selection is different. In ML, any attributes can be considered as potential predictors, potentially suggesting new potential risk factors for diseases like ACS. In this study, all attributes in the registries that fall under the scope of the research were considered as potential predictors and further evaluated using a comprehensive ML feature selection method to improve the prediction performance of the simplest model possible.

The huge and intricate registries dataset requires more advanced analytics and massive data technologies compared to the standard statistical approach. This is because the correlations between the attributes and the outcomes may be complex and multivariate, which may violate linearity assumptions in a statistical model. Thus, ML algorithms present a new opportunity to build clinical prediction models, in general. Hence, DM using ML algorithms is a potential alternative approach to the classic statistical method. The two EHR-based registry datasets are the assets of this study to practically presents the need and potential of ML in prediction modelling.

ML has been used in a limited way to develop prediction models. The earliest study to use ML to construct ACS prediction models was mainly focus on a specific ML algorithm. In fact, one particular novel technique, which has been commonly researched, particularly in developing ACS prediction models, is ANN (Baxt et al., 2002;Harrison and Kennedy, 2005;Green et al., 2006b;McCullough et al., 2007;Bassan et al., 2004). Even though ANN is an intricate technique, it excels at detecting complex and nonlinear correlations. The technique mimics human brain interactions in

processing and understanding relationships. In addition to ANN, the combined method of genetic algorithm (GA) and fuzzy rule has been utilized in developing a UA risk assessment tool(Dong et al., 2014), the decision tree (c4.5) was applied in assessing risk factors of CHD(Karaolis et al., 2010), and the Classification and Regression Tree (CART) was employed to develop mortality prediction models for patients with acute decompensated heart failure (ADHF)(Fonarow et al., 2005). However, there are a wide range of ML algorithms which can be explored in building prediction models, as there is no one specific ML algorithm that best suits all datasets(King et al., 1995, Ali and Smith, 2006). Thus, this study explores a wide range of ML techniques suitable for developing ACS prediction models.

Limited number of studies has been found comparing ML algorithms in developing ACS prediction model. For example, VanHouten et al. (2014) has assessed Random Forest (RF), elastic net, and ridge regression algorithms for building ACS prediction models; Hu et al.(2016) evaluated Support Vector Machine (SVM), RF, Naive Bayes (NB) and LG models for the same task, and Sladojević et al.(2015) compared seven ML algorithms for building an ACS prediction model. These comparison studies were small and based on a limited number of ML algorithms. Considering larger comparisons of ML algorithms, Kurz et al. (2009) tested several ML algorithms available in Waikato Environment for Knowledge Analysis (WEKA) to generate the best ACS prediction models. However, the models that Kurz et al. (2009) developed can only be used for categorical predictors. The issue with models that use only categorical predictors is that they potentially lead to `loss of information.' Attributes, such as heart rate and systolic blood pressure (SBP), need to be discretized to a subset of categorical values, which generally leads to a loss of information from the original data. For instance, within an interval, two numerical values may be at different extremes, but, because they both fall within the interval, the two values are considered equal. Due to this, information is lost which eventually effects the predictive performance of a prediction model. On the other hand, this thesis includes all ACS- type patients and was able to accept both numerical and categorical types of predictors. Furthermore, Kurz et al. (2009) only studied one registry, but our study has two registries

to work with simultaneously that will potentially enrich the discussion on developing ACS models using ML from two different regions with different patient characteristics. It will also inform a discussion on how the characteristics of different datasets could affect the prediction capabilities of the different ML algorithms used.

Most of the models were found to exclude high-risk patients, such as patients with a history of stroke (Lee et al., 1995), patients with a history of cerebrovascular disease, patients with high SBP and diastolic blood pressure (DBP) (Morrow et al., 2001), and patients with persistent ST-segment elevation(Boersma et al., 2000). Also, most of the existing models were developed for a specific type of ACS, i.e., either STEMI(Lee et al., 1995, Morrow et al., 2001, Sladojević et al., 2015) or NSTEMI(Antman et al., 2000, Boersma et al., 2000). Thus, this study, which presents comprehensive coverage with all ACS types and no exclusion of high-risk patients, widens the scope of prediction modelling.

In addition, the advent of new and standard treatments, such as the introduction of potent antiplatelet/antithrombotic agents and the establishment of PCI treatment, has become an issue among the present models. Some of these models were developed before the introduction of this new and standard treatments and the impact of this on the models is still unclear(Kurz et al., 2009). As such, the latest cohorts of patients in the registries used in the study accommodate the current gap in the present models. To the knowledge of the author, this is the first time the Malaysian registry has been used to derive an ACS prediction model. Derivation populations for ACS prediction models have been predominantly Western (Antman et al., 2000, Boersma et al., 2000, Dorsch et al., 2001, D'Ascenzo et al., 2012, Kurz et al., 2009, Huynh et al., 2013, VanHouten et al., 2014).The Malaysian and the UK populations used in this study enable geographical comparisons and furnish further insight into model development for different regions.

1.7. Aims and Objectives of The Study

The important of a reliable prediction model to support CVD and ACS prevention, as well as the availability of vast amount of EHRs within healthcare organizations, motivated this particular research. DM-ML techniques can possibly transform the paradigms of building ACS prediction models in parallel to the future revolution of big data in the medical and health industries. In fact, the growing number of EHRs offers an opportunity for DM-ML to further prove its capability as the best computational solution for both classification and prediction model development in the medical field. Hence, exploring prediction models development for ACS via modern ML techniques could potentially enhance the benefits of using DM in the medical field, in general.

This study aims to explore and investigate a practical method for developing prediction models for predicting ACS in-hospital mortality using DM and ML algorithms on registries' datasets. In the context of ML fields, this study also explores several ML optimization strategies suitable for registry datasets to enhance the overall performance of the developed models. The objectives of the study are listed as follows:

Objective 1: To investigate ML methods and techniques suitable for developing ACS prediction models from registry datasets. The study also aims to establish sets of ML algorithms that are not suitable for building the prediction models.

Objective 2: To investigate ML feature selection methods and techniques for building simpler models with improved prediction power. The thesis also evaluates the potency of predictors of existing ACS models to be adapted to other ACS registries data. Finally, to investigate the strength of predictors from different clinical categories in contributing to model development.

Objective 3: To identify the main causes of misclassification when building ACS prediction models using ML, as well as to develop models using ML to predict the misclassified cases. The study focuses on assessing misclassified cases in terms of minority class, overlapping class, outliers, and missing values.

Objective 4: To investigate and evaluate ML optimization strategies to address an imbalanced dataset and missing values. Do these optimization strategies help in improving the overall model performance? As such, the *overlapped-undersampling* method is proposed to handle imbalanced datasets, while the *mean-clustering-imputation* approach is proposed to handle missing values.

1.8. Outline of Chapters

The rest of this thesis is structured as follows:

Chapter 2provides the literature review on the currently available ACS prediction models. This chapter also presents the review of methods and strategies in developing prediction models using DM and ML.

Chapter 3 describes the relevant methods in achieving the outlined objectives.

Chapter 4 describes and compares the characteristics, patient characteristics, patterns of care, and outcomes of the two datasets. It also describes the process of cleansing, preparing, and transforming the raw datasets for model development. The chapter also elucidates the processes and results of reviewing model development using WEKA, inclusive of several strategies that were employed in developing prediction models.

Chapter 5 provides full documentation of identifying the best set of predictors using ML techniques and its model development. The chapter also presents the findings on the potency of the existing set of predictors in terms of being adapted to other ACS registry data and the strength of predictors from different clinical categories in contributing to the model development.

Chapter 6 presents the results obtained from the misclassification analysis of the classification algorithms. The analysis of the misclassification cases was conducted in terms of the distribution of minority classes, overlapping classes, outliers, and missing values. The chapter also introduces the newly developed model used to predict misclassified cases.

Chapter 7 specifies the optimization approaches used in handling imbalanced datasets and missing values. Moreover, this chapter presents the results of the strategies and discusses the effectiveness of these strategies in contributing to the betterment of the developed models.

Chapter 8 reports on the model validation approach for the best models, specific to Malaysian and UK datasets, as well as the generic models that fit both datasets. Internal validation was performed on each of the best models specific to the regions and generic models. In addition, external validation was carried out on the generic models. Briers scores and calibration graphs of these models are presented, compared, and discussed.

Chapter 9 summarizes and discusses the overall findings. Implications of the study, limitations and future directions are also discussed in this concluding chapter.

Chapter 2: Literature Review

This chapter presents the review of existing and researched ACS prediction models, the methods employed in developing clinical prediction models, as well as the challenges and issues that revolve around developing clinical prediction models using DM and ML algorithms.

2.1. ACS Prediction Model

Prediction has always been a skill set among clinicians. Clinicians need this skill set to make decisions about a certain disease or its severity, and the most appropriate therapy and treatment based on the symptoms. These decisions are normally made based on "intuition" or experience from past cases. Nonetheless, quantifying and rationalising the decisions made can only be made possible with rich clinical data and using the appropriate method. Statistics has been the most widely used and viable method with which to build prediction models up until this time. In addition, clinical prediction models are now being applied as prevention strategies to predict the existence of a disease or high-risk patients with a disease, or even to determine the most suitable therapies for a patient based on diagnosis.

Table 1: Summary of several reviewed ACS prediction models

No	Prediction Models	Year of Publication	Study Design	Derivation Cohorts (Year)	Derivation Cohorts (Country)	n	Range of ACS	Predictors	Predicted Outcome	Time to Outcome	Published C-Statistics	Modelling techniques
1	TIMI (Antman et al., 2000)	2000	RCT	1996-1998	10 countries from North America, South America, and Europe	1957	UA, NSTEMI	* Age * 3 risk factors for CAD * Prior coronary stenosis of>=50% * At least 2 angina events in prior 24 hours * ST-segment deviation * Use of aspirin in prior 7 days * Elevated serum cardiac markers	Death, MI or revascularisation	14 days	0.65	Multivariable logistic regression
2	PURSUIT (Boersma et al., 2000)	2000	RCT	1995-1997	28 countries in Western and Eastern Europe, as well as North and South America	9461	UA, NSTEMI	* Age * ST- Segment depression * Heart rate * SBP * Heart failure * Cardiac enzyme * ST-Depression on presentation	Death and MI	30 days	0.81 (death only) 0.67 (death or MI)	Multivariable logistic regression
3	GUSTO-I (Lee et al., 1995)	1995	RCT	1990-1993	1081 hospitals from 15 countries in North America and Europe, in Israel, Australia, and New Zealand	41021	STEMI	* Age * Killip class * SBP * Heart rate * Anterior Infarction * Previous MI	Death	30 days	NA	Logistic regression

4	GRACE In-Hospital (Granger et al., 2003)	2003	MR	1999-2001	14 countries from North and South America, as well as Europe MONICA Project - 26 countries from Europe, North America, and the Western Pacific	11389	STEMI, UA, NSTEMI	* Age * Killip class * SBP * ST-Segment deviation * Cardiac arrest during presentation * Creatinine level * Heart rate * Initial cardiac enzyme	Death	In-hospital	0.83	Multivariable logistic regression
5	EMMACE (Dorsch et al., 2001)	2001	RR	1995	UK	3684	Acute MI	* Age * Heart rate * SBP	Death	30 days	0.79	Multivariable logistic regression
6	SRI (Morrow et al., 2001)	2001	RCT	1997-1999	Western and Eastern Europe, North America and Latin America	13253	STEMI	* Age * Heart rate * SBP	Death	30 days	0.78	multivariable logistic regression
7	C-ACS (Huynh et al., 2013)	2013	RR	1999-2001	Canada	4627	STEMI, NSTE-ACS	* Age * Killip class * SBP * Heart rate	Death	In-hospital/ 30 days	0.75	Multivariable logistic regression/Hosmer-Lemeshow

8	AMIS (Kurz et al., 2009)	2012	RR	1997-2005	Switzerland	7520	UA, NSTEMI, STEMI	* Age * Killip Class * SBP * Heart Rate * Pre-hospital cardiopulmonary resuscitation * History of heart failure * History of cerebrovascular disease	Death	In-hospital and 12 months	0.875	WEKA - Average One-Dependence Estimators (AODE)
9	SD (VanHouten et al., 2014)	2014	EHR	2007 - 2012	USA	20078	ACS, NON ACS	88 predictors	ACS patients	NA	0.848	Random forest
10	Serbia (Sladojević et al., 2015)	2015	EHR	2008-2011	Serbia	2030	STEMI	* Age * SBP * Diastolic blood pressure * Heart rate * Lvef * Troponin value	Death	In-hospital	0.91	WEKA - Alternating Decision Tree (ADT)
11	MACE (Hu et al., 2016)	2016	EHR	NA	China	2930	ACS	268/284 predictors	Adverse cardiovascular events	NA	0.724	Random forest

Table 1 provides a summary of the reviewed ACS prediction models. As presented in the modelling techniques column (last column), 7 out of 11 reviewed models were developed using a statistical method, i.e., LG. As explained in Section 1.6, when using this conventional method, the candidate predictors are generally pre-selected based on clinical expert opinions, known risk factors, findings from clinical studies, and/or previously developed models. The actual predictors used to developed the model are selected by running univariate and multivariate LG to identify the significance of the candidate predictors. The approach contrasts with the ML paradigm since, in ML, any attributes can be considered as candidate predictors, which allows for identification of potentially new potent predictors which can subsequently suggest new potential risks for ACS.

In a large dataset, the assumption of data linearity using the conventional statistical modelling approach may be violated due to complex and multivariate correlations between the attributes and the outcome. More advanced analytics and massive data technologies are required to handle the large and intricate data in a registry or EHRs. Thus, DL-ML presents a new advancement in predictive model development compared to the statistical approach. Table 1 demonstrates that models developed using ML techniques such as AMIS, SD, and Serbia have better predictive power than models developed using the statistical approach (TIMI, PURSUIT, GUSTO-I, GRACE, EMMACE, SRI, C-ACS).

ML has been used in a limited way to develop prediction models. The earliest study to use ML to construct ACS prediction models was mainly focus on a specific ML algorithm. In fact, one particular novel technique, which has been commonly researched, particularly in developing ACS prediction models, is ANN (Baxt et al., 2002;Harrison and Kennedy, 2005;Green et al., 2006b;McCullough et al., 2007;Bassan et al., 2004). In Table 1,the three reviewed ACS models that were developed using ML were compared to a limited number of ML techniques i.e., the SD - 1 ML algorithm, Serbia - 7 ML algorithms and MACE - 4 ML algorithms. And, only the AMIS model was compared to broad number of ML algorithms from WEKA. However, the AMIS model was only able to handle categorical

predictors. As mentioned previously, the potential for information loss is critical for those predictors that are originally numerical, such as SBP, DBP, heart rate, and creatinine level, and this, in turn, eventually affects the predictive performance of a model.

Different sets of predictors are observed for each of the models in the reviewed ACS models. Each set of predictors seems to fit with the sample populations that the model was derived from, and most of the models were derived from a particular region, with the exception of TIMI, PURSUIT, GUSTO-I, GRACE and SRI. Regardless of the varied clinical outcome of the models, the most common predictors incorporated in ACS models are age, heart rate, SBP, and killip class, which are also the key risk factors for ACS and CVD, in general. The potential predictors for ACS prediction models are generally derived based on the availability of information on the point in clinical cause, in which the model can be used for prediction or risk assessment. For example, upon admission, any information obtained refers to demographic type, such as age, gender, and ethnicity, whereas basic clinical presentation denotes heart rate, SBP, height, and weight. At this point, normally, both basic physical examination and typically occurring symptoms are assessed. In addition, the medical history of patients and their basic health lifestyle, such as smoking status, is also recorded at this point. Examples of predictors under this category that have been applied in the models of Table 1 are history of heart failure, diabetics, and history of cerebrovascular disease. Upon completion of an ECG, its related parameters, such as ST-segment, Q-wave, and T-wave, become another set of information that is made available at the first stage of ACS clinical cause. In fact, prediction models for use in an emergency department and some models built with the objective of early-risk stratification, such as the C-ACS and AMIS models, mainly utilize such first-hand information as their predictors (Green et al., 2006b).Other than that, the results of biomarker tests offer more useful information in terms of distinguishing between the three ACS categories. Depending on the symptoms and ECG status, biomarker tests, such as cardiac troponin and MB fraction of creatinine kinase (CK-MB) measurements, usually take place at a later stage(Bassand et al., 2007). GRACE and SD are examples of models that considered

biomarker data as predictors. Although biomarker tests contain useful information with which to identify different categories of ACS, they may not be popular with those simple models with the need for early stratification, such as AMIS, EMMACE, SRI, and C-ACS.

DM-ML can also be utilized to identify essential predictors in relation to the outcome known as feature selection. Feature selection aids in choosing a set of predictors that improves the predictive power with less data. It can identify nonessential, irrelevant, and redundant attributes that may affect the accuracy of the model. A model with cost effective predictors is preferable due to its simplicity, and ease of understanding and explanation(Guyon and Elisseeff, 2003). Feature selection, when applied, needs to be embedded as part of the model selection process. Models applying feature selection or different set feature selection methods and models without feature selection should be compared and analysed. Better predictive power indicates the best set of predictors for the model. For example, Fonarow et al. (2005)developed a model for risk stratification of in-hospital mortality for acute decompensated heart failure (ADHF). They employed CART to establish the best predictors for the model, in addition to constructing a risk stratification model. Karaolis et al.(2010) evaluated a decision tree (DT) model in terms of identifying the significant risk factors for myocardial infarction (MI), PCI, and CABG. The study claimed the promising correctly classified rate, indicating that the identified risk factors were the most important. They found that age, smoking, and history of hypertension were important risk factors for MI; family history, history of hypertension, and history of diabetes were important risk factors for PCI and CABG; and age, history of hypertension, and smoking were important risk factors for CABG. In addition, Vinterbo and Ohno-Machado (1999)built a model based on LG and identified a set of predictors using a genetic algorithm. By comparing the AUC results, the study suggested that a genetic algorithm was significantly better than the standard backward, forward, and stepwise variable selection methods.

From Table 1, the implementation of ML feature selection of models developed using ML as modelling technique (AMIS, SD, Serbia and MACE) was further reviewed. Even though the Serbia model was developed using a

ML technique, the candidate predictors of the model are pre-selected, then further reduced using ML techniques. The AMIS model, on the other hand, uses the common feature selection method, i.e., the sequential backward deletion method, to determine the best predictors for the model. Whilst, the SD and MACE models do not apply any feature selection method in order to make the model simpler. However, the MACE model is the only model from the reviewed list that retrieves its predictors from free text admission records. The extraction of predictors is implemented by combining a rule-based approach with an ML method known as Conditional Random Fields.

Most of the reviewed ACS models were derived either from RCT, registry, or EHR data. RCT, which follows rigorous scientific principles, is considered the most powerful tool in clinical research. The samples collected are considered more accurate and complete with defined prospective , methods, and duration. However, the emergence of EHRs in medical field has promoted the use of registry or EHR data to derive the data for model development. The registry or EHR-based data fills the gaps in RCT, such as not representing real populations and the potential of bias due to defined inclusion and exclusion criteria. Also, a registry based on EHR data is considered essential evidence in acquiring the best external evidence in evidence-based medical applications. Thus, many ACS models now derived their data from EHR-based registries, as in GRACE, EMMACE, C-ACS, AMIS, SD, Serbia, and MACE models.

Furthermore, most cohorts of the reviewed ACS models in Table 1 were derived from Western populations. TIMI, PURSUIT, GUSTO-I, and SRI are examples of models derived from mainly Western countries. Other models, such as EMMACE, C-ACS, AMIS, SD, and Serbia were derived from a specific Western country. The use of Asian populations has been limited in deriving ACS model. Among the 11 models of Table 1, MACE is the only model that was built based on Asian cohorts i.e. China. However, some of cohorts included in GRACE comes from Western Pacific that might includes some patients from Asian regions.

Mortality among patients with chest pain at varied time points is a common outcome in ACS prediction models. Nine out of 11 reviewed

models of Table 1 have mortality as the outcome. Besides than mortality outcome, TIMI and PURSUIT also have MI as the end point. SD, on the other hand, distinguishes between ACS and non-ACS patients as the outcome of the model, while MACE distinguishes adverse cardiovascular events.

In Table 1, the ACS models were mainly for selective patients with either STEMI or NSTEMI/UA. Only GRACE, AMIS, and MACE models cover all the ACS spectrum. There were also issues of exclusions of high risk patients in some of these models. For example, the GUSTO-I model excludes patients with history of stroke, the SRI model excludes patients with high SBP and DBP, and the PURSUIT model excludes patients with persistence ST-segment elevation.

Among all of these models, the most popular ACS prediction models are TIMI and GRACE, as they have been validated by many other populations. The following further summarizes each of the reviewed ACS models.

Thrombolysis in Myocardial Infarction (TIMI)

The TIMI prediction score appears to be the most widely used model, and it is known for its use for patients with NSTEMI and UA. This prediction, or risk, score was derived from 1957 patients from the TIMI 11B trial, which consisted of cohorts from 10 countries, including North America, South America, and Europe, in which the samples were given unfractionated heparin for the study(Antman et al., 1999). The TIMI risk score predicts mortality, new MI, and urgent need of revascularization within 14 days of the event by summing the scores of 7 predictors (Antman et al., 2000). The 7 predictors are: age greater than 65 years old, at least three risk factors of CAD, prior coronary stenosis of 50% or more, ST Deviation on ECG, two angina events that occur within 24 hours, use of aspirin in the past 7 days, and elevated serum cardiac markers.

The risk score was built by using multivariable LG and was validated with three different groups:1) an enoxaparin group from TIMI11B (n=1953), 2) an unfractionated heparin group from the ESSENCE trial(n=1564), and 3) an enoxaparin group from the ESSENCE trial (n=1607). The c-statistic for

the derivation cohorts was 0.65 and fell within the range between 0.65 and0.59 for the validation cohorts. Furthermore, the simplicity of the risk model has turned it into the present guidelines for admission and treatment decisions(Anderson et al., 2007).

The TIMI risk score was also used as the platform to generate risk score for STEMI patients(Morrow et al., 2000). In fact, many had validated TIMI risk score in various cohorts, end points, and all types of ACS(D'Ascenzo et al., 2012). For instance, TIMI risk score was validated on 30 days and 1 year mortality of patients suffering from myocardial revascularization during their initial hospitalisation (de Araújo Gonçalves et al., 2005). Meanwhile, Morris et al.(2006), Chase et al. (2006)and Hess et al.(2010)validated TIMI risk score with patients who had chest pain at the emergency department. As a result, the modified TIMI risk score with extra weight on ischemic ECG changes and tropinin elevations displayed better risk stratification among the patients with chest pain at the emergency department (Body et al., 2009).

The Receptor Suppression Using Integrilin Therapy (PURSUIT)

PURSUIT is another risk model for UA and NSTEMI. This model was derived from 9461 patientsin28 countries in Europe, and North and South America (Boersma et al., 2000). The end point of this risk model is mortality or mortality and MI of UA and NSTEMI patients within 30 days. The predictors are age, ST-Segment depression, higher heart rate, lower SBP, signs of heart failure, and cardiac enzyme upon admission. Unlike the TIMI score, PURSUIT presents a more complex calculation. For instance, age has a range of 0-6 scores depending on various age ranges to predict mortality. Nonetheless, in a review of validated ACS models performed by D'Ascenzo et al.(2012), only two studies had validated the PURSUIT model.

The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO)

The GUSTO risk score was drawn from 1081 hospitals located in 15 countries in North America and Europe, and including Israel, Australia, and New Zealand (Investigators, 1993). The GUSTO trial was a randomized clinical study involving MI patients eligible for fibrinolysis (Lee et al., 1995).

The study examined four thrombolytic agents with aspirin and BB. The model predicts1 year and 30 day mortality after MI with the most significant predictors, which are age, SBP, heart rate, infract location, and prior MI conditions. The GUSTO risk model has been validated in predicting three vessel diseases together with the TIMI, GRACE, and PURSUIT models. The GUSTO risk model achieved an AUC of 0.63, lower than the AUC scores achieved by TIMI (0.71), GRACE (0.68), and PURSUIT (0.65) (Isilak et al., 2012). The GUSTO model has also been used to validate the effect of medications, such as use of atenolol for acute MI after thrombolysis (Pfisterer et al., 1998).

The Global Registry of Acute Coronary Events (GRACE)

The GRACE model is a widely-known ACS model derived from a patients' registry (n= 11 389) with a complete range of ACS(Granger et al., 2003). In addition, the registry was comprised of unselected patients from North and South America, Europe, and the Western Pacific region. The primary function of the model is to predict in-hospital mortality. Furthermore, the model was built using multivariable LG and was presented via an intelligent scoring system. In comparison to the TIMI and PURSUIT scoring systems, the GRACE risk score presents a more complex calculation with a detailed gradation for each predictor. The predictors for the GRACE model are age, killip class, SBP, ST-Segment deviation, cardiac arrest during admission, serum creatinine level, heart rate, and initial cardiac enzyme level. Moreover, the model has achieved a rather excellent c-statistic of 0.83. The model was further validated in two varied cohorts with c-statistics of 0.84 and 0.79.

In addition, when compared to the TIMI score, the GRACE model emerges as one of the most validated ACS models. Furthermore, D'Ascenzo et al. (2012) claimed that the GRACE had been validated in 12 studies within multiple clinical settings with a total of 36,517 patients, and that the average AUC of the GRACE model in validation studies that consisted of ACS or UA/NSTEMI cohorts for both short- and long-term outcomes had been 0.85. Due to its exceptional prediction ability in varied clinical settings, the

European Society of Cardiology has recommended GRACE as a suitable risk stratification model for NSTEMI patients(Bassand et al., 2008).

Evaluation of Methods and Management of Acute Coronary Events (EMMACE)

According to Dorsch et al.(2001), the EMMACE model appears to be one of the simplest ACS prediction models with only three predictors, which are age, SBP, and heart rate. In fact, EMMACE was derived from registries gathered in the Yorkshire region in the UK that predicted 30 day mortality among patients with acute MI. The model achieved an AUC of 0.79, while an AUC of 0.76 was achieved by the tested cohorts. EMMACE has been validated over a wider ACS diagnosis and maintained its c-statistics of 0.77 to 0.78(Gale et al., 2008b).

Simple Risk Index (SRI)

Apart from the EMMACE model, the SRI model is another example of a simple and rapid risk score model that predicts 30 day mortality among STEMI patients(Morrow et al., 2001). With predictors similar to those of the EMMACE model, the SRI model was derived from 800 hospitals found among Western and Eastern Europe, North America, and Latin America (Giugliano et al., 2001). In fact, the c-statistics of the derivation model for 30 day mortality have been 0.78 and 0.79 when validated on external cohorts.

The Canada Acute Coronary Syndrome Risk Score (C-ACS)

C-ACS model was aimed to serve early risk stratification among patients with ACS. The risk model was derived from two Canadian ACS-1 (C-ACS) registries(Huynh et al., 2013). C-ACS was a prospective study for STEMI and NSTE-ACS patients. Additionally, it was claimed that the model was simple and had no need of a system or calculator to estimate in-hospital or 1 year mortality. It uses only four predictors, i.e., age, killip class, SBP, and heart rate. However, all the predictors are categorical, which might result in loss of information, and thus could affect the reliability of the performance. The model was also validated on four other datasets, with an average c-statistic of 0.75, for short-term mortality.

Acute Coronary Myocardial Infarction in Switzerland(AMIS)

The AMIS model was built by using the ML algorithm known as "Averaged One-Dependence Estimators (AODE)" to predict in-hospital mortality of 7520 patients with ACS obtained from the AMIS-Plus registries of 2001-2005. In the study, several other ML algorithms available in WEKA were tested, and AODE emerged as the best. Through the use of minimal features available at first patient contact, i.e., age, killip class, SBP, heart rate, pre-hospital cardiopulmonary resuscitation, history of heart failure, and history of cerebrovascular disease; a c-statistic value of 0.875was attained. The model was externally validated on the Krakow cohorts, achieving a c-statistic of 0.842 (Kurz et al., 2009).

Vanderbilt University Medical Centre's Synthetic Derivative (SD)

VanHouten et al. (2014)developed ACS prediction models by using two ML algorithms, which were elastic net and RF, with 20,038 suspected ACS patients from EMRs. The aim of the model was to provide an automated prediction model where a new prediction is calculated as new data is entered into the EMR system. In the study, all noisy data and missing values were embedded in the developed dataset. The missing values were imputed with median value. As a result, the best model with an AUC of 0.848 was built using RF with 88 predictors. However, the model, although it has been internally validated, has never been validated with new data.

The Serbia

This ACS model was derived from a cohort in the information system of the Institute for Cardiovascular Diseases of Vojvodina, Sremska Kamenica, Serbia (Sladojević et al., 2015). Models were developed by using seven ML algorithms from WEKA. As a result, Alternating Decision Tree (ADT) with a cost sensitive classification was found to be the best model for estimating mortality among STEMI patients who underwent PCI. Cost sensitive classification allows for the re-weighting of instances to reflect the defined misclassification cost(Witten et al., 2005). The final model was constructed based on 6 predictors. New data from the same population was

used to validate the model. The model maintained a good performance, with an AUC 0.82, when validated with new data.

Major Adverse Cardiovascular Event (MACE)

Hu et al. (2016)employed a dataset derived from 2930 unstructured admission records from a Chinese hospital. Information concerning potential predictors was extracted using rule-based medical language processing (RBMLP) and CRF, which had been used to develop the models based on four various ML algorithms (Hu et al., 2016). As a result, the RF emerged as the best model with an AUC of 0.724. As future research Hu et al. (2016) aim to validate the model on a large scale with an EHR dataset of different cohorts.

2.1.1 The study

This study is motivated by the belief that DM-ML is able to build a better model in terms of predictive power (c-statistic) compared to statistical methods. Although TIMI and GRACE (both models were developed using a statistical method) are the most validated and accepted as present clinical guidelines, the author believes that DM-ML could produce a better discriminatory performance than a statistical method could. Further, by presenting a practical way of developing prediction models with application of other ML techniques such as feature selection method and ML optimization strategies , more validations could be initiated from this study to promote models using DM-ML for acceptance in clinical practise and preparation of big-data evolution in medical field.

Models developed using ML techniques are limited, and most of them only compared on limited number of ML techniques . Although the AMIS model compared a broad range of ML algorithms, the model can only handle categorical predictors, which may, as mentioned previously, affect the performance of the model. Thus, this study develops ACS prediction from 29 ML algorithms that are able to handle numerical and categorical predictors from two EHR-based registries from Malaysia and Leeds, UK. The different region of the registries provide more insight into the different characteristics of two registries and the effect on different ML algorithms.

In addition, the study also evaluated a number of ML feature selections methods as a way to improve the predictive performance of the model, as well as to make the model simpler. None of the reviewed models in Table 1 explored ML feature selection methods as a way to improve the performance of the model. In addition, this study also investigates the reason for the degradation in performance of models developed using ML techniques.

The registry datasets used in the study include patients with all three types of ACS, patients with new treatment, and do not exclude high-risk patients. Furthermore, the registry datasets used in the study contain the challenges found in EHR-based registries and practically present how DM-ML methods could use strategies for handling the data quality issue in a registry dataset, as well as improving the predictive power of the prediction model.

2.2. Medical DM Challenges and DM-ML Optimization Strategies

2.2.1 Complexity of Medical Data

The nature of medical data is challenging for DM. According to Fayyad et al. (1996), challenges may arise due to the nature of data and the granularities of knowledge to be extracted. The intricacy of medical data mainly originates from the biological and social complexities of a patient (Beale, 2005). Moreover, the growth of data is extremely rapid and sizeable. A patient admitted to an intensive care unit (ICU) may have 50 or more parameters collected per hour. Heterogeneity, which lies in different sources of data, different kinds of data, and data originating from different systems, contributes to data complexity. Such data may originate from doctors, clinicians, or even health administrators(Hayrinen et al., 2008). Medical data are captured for varied purposes. Inputs for diagnosis, prognosis, and treatment have their own purposes and meaning in a medical dataset (Hayrinen et al., 2008). Additionally, different types of data are captured in a medical database, ranging from numerical values, images, sounds, to unstructured free texts (Cios and Moore, 2002). Sounds and free text values can easily be ambiguous, inconsistent, and vague. Thus, medical data,

which have no standard or formal structure, render further challenges in medical DM.

Therefore, it is vital to first comprehend the context of the domain and the cohorts involved in the study. The attributes of the patients involved in the study, as well as the inclusion and exclusion criteria, need to be properly defined and reported. The context of the dataset source has to be understood and should match with the objectives of the prediction model. Furthermore, target outcome, potential predictors, and target user of the prediction models are some essential aspects that have to be understood and described. With that, understanding the characteristics of a dataset related to a prediction model should also suggest some suitable DM techniques and approaches.

2.2.2 Feature Selection

The complexity and rapid growth of medical data have increased the dimensionality of the data, thus resulting in irrelevant, redundant, and noisy attributes. In DM, a model built from a large number of attributes ("curse of dimensionality") may possibly have deteriorated predictive power(Nisbet et al., 2009). In fact, the term "curse of dimensionality," coined by Richard Bellman (1961), refers to the issue of data when they become sparse as the volume of data increases, hence causing inefficient predictive power. However, one way to reduce the dimensionality of a dataset in DM is to reduce the number of features or attributes to a manageable number without jeopardizing its DM objectives. Moreover, prior studies have shown that the selection of a significant set of attributes facilitates data visualization, data understanding, reduces overfitting, and improves the overall prediction performance (Das, 2001, Saeys et al., 2007, Guyon and Elisseeff, 2003, Tan, 2007, Hall and Smith, 1998, Blum and Langley, 1997, Kotsiantis et al., 2007). In addition, upon reducing the attributes, a faster training time and simpler model can be attained. Rapid training time is an important consideration when dealing with a huge dataset, while a simpler model offers deeper insight into the underlying processes that generate the data.

Within the context of classification modelling, the feature selection method can be classified into three categories: 1) the *filter* method, 2) the

wrapper method, and 3) the hybrid and embedded method. The *filter* method ranks its features by evaluating an individual feature (*univariate-filter*) or by evaluating an entire subset of features (*multivariate-filter/subset*). Some properties used in evaluating the features are information (e.g., information gain), distance (e.g., Euclidean distance), consistency, similarity, and statistical (e.g., Chi-square) (Jović et al., 2015). As for the multivariate-filter method, apart from evaluating features, the selection of a subset of feature relies on the search strategy for the set. Commonly, the search proceeds in one of the following ways: 1) it starts with the empty set and features are added into the set (forward selection), 2)it starts with a full set of features and some are eliminated from the set (backward elimination)3) it starts with the empty set and a full set of features, simultaneously from both dimensions (bidirectional selection),or 4) it uses a genetic algorithm to identify the set of features (heuristic feature subset selection).

The main advantage of the filter method is that it does not depend on any classification algorithm. Hence, the method is simple, fast, and can easily scale to a high-dimensional dataset. However, in the *univariate-filter* method, the focus of evaluation is only on an individual feature and the outcome, thus dismissing its correlations with other features. Hence, some important information gained by forming a combination of features that could generate a better model may be disregarded(Guyon and Elisseeff, 2003, Blum and Langley, 1997).

An example of the *univariate-filter* method is Information Gain. Information Gain measures the relevancy of each individual feature towards its outcome, while ignoring correlations with other features. In addition, *Correlation-Based-Feature-Selection (CFS)* algorithms are an example of the *subset* approach. *CFS* identifies a subset of features via selection with high correlation with the class, but low correlation with each feature(Hall and Smith, 1998). Karegowda et al.(2010) stated that the *CFS* method was better than the Gain ratio in comparative studies on varied domains. In addition, Zhang et al.(2008a) proposed a new filter method named the Constraint Score, which applied pairwise constraint, instead of class label information, when selecting the feature of classification development. As a result, the Constraint Score offered better results compared to the Fisher

Score and Laplacian Score algorithms when tested on high-dimensional datasets(Zhang et al., 2008a). Furthermore, Yin et al. (2012) introduced a new feature selection method that incorporated Hellinger distance to calculate a measure of distribution divergence, which is claimed to be more efficient and effective when dealing with high-dimensional datasets.

Next, the *wrapper* method is a method that incorporates a learning algorithm into evaluating the selection of features (Kohavi and John, 1997). The main advantage of this *wrapper* method derives from the advantages of subset feature selection and the specific classifiers. Nonetheless, compared to the *filter* method, the *wrapper* method has a higher computational cost due to the additional evaluation of the subset of features with the specific learning algorithm. This method also tends to overfit the learning algorithm used to evaluate the subset of features. Hence, the recommendation is to develop the model on other classification algorithm, instead of using the algorithm used for feature selection. For instance, John et al.(1994)wrapped around a subset selection with an induction algorithm to consider its biasness, in addition to defining two levels of relevancy in selecting a subset of features. Additionally, Maldonado and Weber (2009)introduced a *wrapped* method using a SVM algorithm, which utilizes a sequential backward elimination method to remove insignificant features and applies a random split in each iteration of subset feature selection. Therefore, the method presents better results than other *filter* and *wrapper* methods. The method avoids overfitting and is flexible to any kernel function (Maldonado and Weber,2009). Earlier, Weston et al.(2001)introduced a feature selection method on an SVM algorithm that is applicable to a non-linear kernel.

A hybrid method, on the other hand, combines both *filter* and *wrapper* methods. The *filter* method identifies the potential subset features, while the *wrapper* method determines the best subset features. For example, Yang et al. (2010) implemented the information gain approach to identify a potential of subsets features, and later used the *wrapper* method with a genetic algorithm to identify the best subset features in selecting relevant genes from a microarray dataset. Meanwhile, Bermejo et al.(2012) improved the hybrid feature selection method by reducing the use of the *wrapper* method in selecting the best subset features. Beyond this, the embedded method

embeds feature selection into the classification algorithm, hence becoming part of the model construction. This feature selection method, however, depends on the classifier used. The DT filter was used to weigh the attributes in the new weighted NB algorithm introduced by Hall (2007), while Guyon et al. (2002) incorporated Recursive Feature Elimination (RFE) in the SVM algorithm to select the best set of genes for cancer classification.

Hence, feature selection offers great utility in finding potential predictors in developing a prediction model. In identifying the best predictors for the prediction model, A., Sudha et al. (2012)employed a subset feature selection algorithm, Huang et al. (2004)used the ReliefF algorithm, Kurz et al. (2009) applied sequential the backward deletion method, and Khosla et al. (2010)proposed a novel feature selection algorithm called Conservative Mean feature selection.

2.2.3 Missing Data

Medical databases, in particular, are dense with missing values(Cios and Moore, 2002). Missing data can be completely at random (MCAR), missing at random (MAR),or missing not random (MNAR). MCAR means that there is no obvious pattern to the missing values of the observed data, for instance, when a clinician unintentionally fails to record the height of a patient. As for MAR, the missing values can be observed in a particular subsample or several subsamples, but no missing pattern appears in the entire sample, such as information on patients receiving PCI treatment being blank for those diagnosed with UA/NSTEMI. In a standard guideline, PCI treatment is a specific treatment for STEMI. Hence, both MCAR and MAR are known as ignorable patterns since any model explaining the missingness can be ignored, and the outcome from the analysis is still valid. Even though a model for missingness can be ignored, appropriate measures, such as excluding missing data from the sample or applying an imputation strategy, should be implemented to improve model performance(Pedersen et al., 2017). MNAR, on the other hand, refers to missingness that depends on the missing values or other unobserved predictors. Hence, for missingness due to MCAR, a careful analysis has to be performed so as to understand why the data are missing and the probable values. A model of the missing

values must be part of the process of inference to avoid bias. This review focuses on strategies for addressing MCAR and MAR.

One simple way of handling missing data of the MCAR or MAR is to exclude all cases or all features with missing values(if there are many missing values of the said feature). However, this would reduce the total number of observations in the training set or remove important features that may happen to be crucial predictors. As such, Delen et al. (2010) took an approach of excluding features with 95% missing values from the training set, as they were insignificant to the prediction model. Another way to handle missing data is to impute a significant value, such as the mean or the common value of the categorical type (Green et al., 2006a, Dangare and Apte, 2012, Khosla et al., 2010). Meanwhile, Khosla et al.(2010) employed the Linear Regression and Regularized Expectation Maximization methods for data imputation. Hruschka et al.(2004)proposed an approach using clustering in estimating the imputation values. In his study, the approach was to first cluster the complete instances by class label, and, then, impute the mean of the nearest cluster for each instance that contains missing values. In addition, Zhang et al.(2008b) also used the cluster-based approach to propose an imputation strategy. The study proposed a kernel function nonparametric random imputation to estimate the imputation value of each cluster, by which the training samples were then clustered using the K-Means algorithm and ignoring the class label. Apart from that, Grzymala-Busse and Hu(2001) compared nine methods of handling missing data, which finally concluded that the C4.5 method, based on entropy and splitting, and excluding the missing attributes emerged as the two superior methods for addressing missing data. Reviewing several practical imputation methods on risk modelling using real clinical datasets with binary outcomes, Ambler et al. (2007) concluded that models developed by ignoring and using only complete instances potentially produce unreliable models with substantial bias. Also, the study suggested that multiple imputation by chained equations (MICE) was the best multiple imputation method, and that conditional imputation worked well on those datasets with the same characteristics as their dataset.

In addition, some ML algorithm can also handle missing values by itself, but there is no specific strategy for specific missing patterns (MCAR, MAR, or MNAR). As such, Su et al. (2008) reviewed the methods used in handling missing values by each of ML algorithm in WEKA. These can be categorized into: 1) ignoring missing values - as applied in NB, DT, Projective Adaptive Resonance Theory(PART) algorithms2) imputing missing values with the mean or median of the observed values - as applied in LG and RF3) replacing missing values with a default value - as applied in SVM and One Rule (OneR) 4) imputing missing values with a distance measure - as applied in K-Nearest Neighbour (KNN) and 5) having a specific algorithm to handle missing values- as applied in ANN.

2.2.4 Imbalanced Dataset

An imbalanced or skewed dataset is defined as having an uneven distribution between classes, where one class has a lower distribution than the other. An imbalanced class is indeed a concern in DM(Yang and Wu, 2006). The imbalanced distribution of the dataset may create biased results. In fact, most ML algorithms, such as DT, tend to predict based on the majority class data; hence, this results in a higher probability of misclassification of the minority class. Unfortunately, an imbalanced class distribution reflects many real-world situations, such as fraud detection in banking transactions, facial recognition, and oil spill detection. Similar scenarios also exist in medical DM, such as identification of a particular disease, and prediction of mortality among patients suffering from breast cancer.

A number of studies have suggested many approaches for handling imbalanced classes, which can be categorized into two groups: 1)data- level approaches, and 2) algorithm-level approaches. A data-level approach balances the distribution of classes at the processing level (re-sampling), and reduces the effect of skewed data during the learning process. Hence, there is no dependence on a learning algorithm. In general, a balanced distribution can be achieved either by reducing the majority class (*undersampling*) or by increasing the distribution of the minority class (*oversampling*). Meanwhile, an algorithm-level approach involves adjusting the decision threshold, specifying costs for each class, enhancing the

sensitivity of the existing algorithm towards the minority class, combining multiple algorithms for classification, and constructing a new algorithm that works better on imbalanced datasets. Boosting and bagging are examples of algorithm-level approaches.

Due to a high number of negative classes, Barakat et al.(2010) used a K-Means algorithm as a sub-sampling method to select a dataset for training. In addition, Rahman and Davis(2013) suggested a cluster-based oversampling method, where the majority and minority classes were clustered using K-Means clustering. Next, the different sets of clusters for both the majority and minority classes were combined and learned by DT and FURIA classifiers. The results showed that the whole set of minority classes and a cluster of majority classes, as well as two clusters from the minority class and a cluster from the majority class, improved the overall sensitivity and specificity of the models. In addition, this method is better compared to the cluster-based undersampling approach proposed by Yen and Lee(2009).

Meanwhile, the Synthetic Minority Over-Sampling Technique (SMOTE) method generates synthetic instances for the minority class by learning from examples in the minority class. The synthetic instances were included in the training set to build the classification model (Chawla et al., 2002). Additionally, an analysis performed by Han et al.(2005) showed that the data points were far from the borderline with minimum impact upon classification. Thus, the study suggested an approach to building synthetic instances of the minority class based on the examples of data points near the borderline. This approach is known as borderline-SMOTE. Moreover, Ramentol et al. (2012) used SMOTE to create additional samples for the minority class, and utilized the Rough Set Theory to improve the quality of the minority samples created by SMOTE.

Due to the advantages offered by under-sampling and over-sampling methods, Batista et al.(2004) combined both methods in solving the issue of imbalanced dataset. Meanwhile, Kubat and Matwin(1997) removed instances labelled as noise from the majority class to balance the class distribution. Moreover, Khalilia et al. (2011)implemented a repeated sub-

sampling method based on ensemble learning to handle an imbalanced dataset. In addition, Stefanowski (2013) revealed that the small number in the minority class was not really due to performance degradation. Instead, the problem lies in a minority class that consists of small sub-parts and overlapping borderlines between classes. Hence, in tackling imbalance issue due to borderline examples in minority classes, the study found that the under-sampling method using Neighbourhood Cleaning Rule(NCR) and hybrid SPIDER techniques (framework that integrates a selective data pre-processing with the ensemble method) displayed better results in comparison to the oversampling method using SMOTE and the one-sided method.

Boosting refers to a technique that iteratively increases the weight of misclassified instances and lowers the weight of correctly classified instances(Freund and Schapire, 1996). Apart from reducing overfitting, the bagging technique can also handle an imbalanced dataset. In the bagging technique, various training samples are generated by replacement (Bauer and Kohavi, 1999). Meanwhile, Galar et al. (2012)reviewed the capability of varied ensemble methods in handling imbalanced datasets. The combination of undersampling technique with a bagging ensemble algorithm led to positive results.

Moreover, Japkowicz et al.(2002) concluded that the imbalanced dataset issue is more severe when the training set is small, which would imply a greater impact on classifiers sensitive to imbalanced datasets. The study also found that compared to multilayer perceptron (MLP) and SVM, the C4.5 algorithm exhibited higher sensitivity towards imbalanced datasets. Additionally, the study also found that a cost-modifying method was better at handling an imbalanced dataset than undersampling and oversampling methods. Yin et al.(2012), on the other hand, focused on an imbalanced issue at the feature selection phase. As such, two methods were introduced in feature selection for classification of imbalanced datasets: 1) a decomposition-based framework for any existing feature selection method that can be embedded into the framework, and 2)the Hellinger distance-based methods. The work showed that the set of features derived from these two methods led to a better classification model compared to three

existing feature selection methods: the *CFS*, Fisher, and Mutual Information methods. In addition, Japkowicz(2000), Chawla(2010), and Kotsiantis et al.(2006) reviewed different strategies and algorithms in handling imbalance datasets.

In dealing with an imbalanced dataset, a more appropriate evaluation technique is required. Evaluation based solely on the accuracy rate does not present true classification results. For example, in an extremely imbalanced dataset, the accuracy may still be good, even though all the correctly classified cases are from the majority class and none are from the minority class. Besides, the accuracy rate does not differentiate the correct classification of majority and minority classes. In fact, the ROC Curve, AUC, and F-Score are some instances of evaluation metrics that can be employed to evaluate the performance of the model when dealing with imbalanced datasets (Witten et al., 2005, Baldi et al., 2000).

Chapter 3: Research Methodology

This chapter describes the methods employed in the study. The methodology applied in this study was guided by the clinical modelling approach as depicted in Steyerberg(2009) and Lee et al. (2016), along with standard DM methodology, i.e., CRISP-DM (Nisbet et al., 2009). A summary of the methodology used in this research is portrayed in Figure 1.

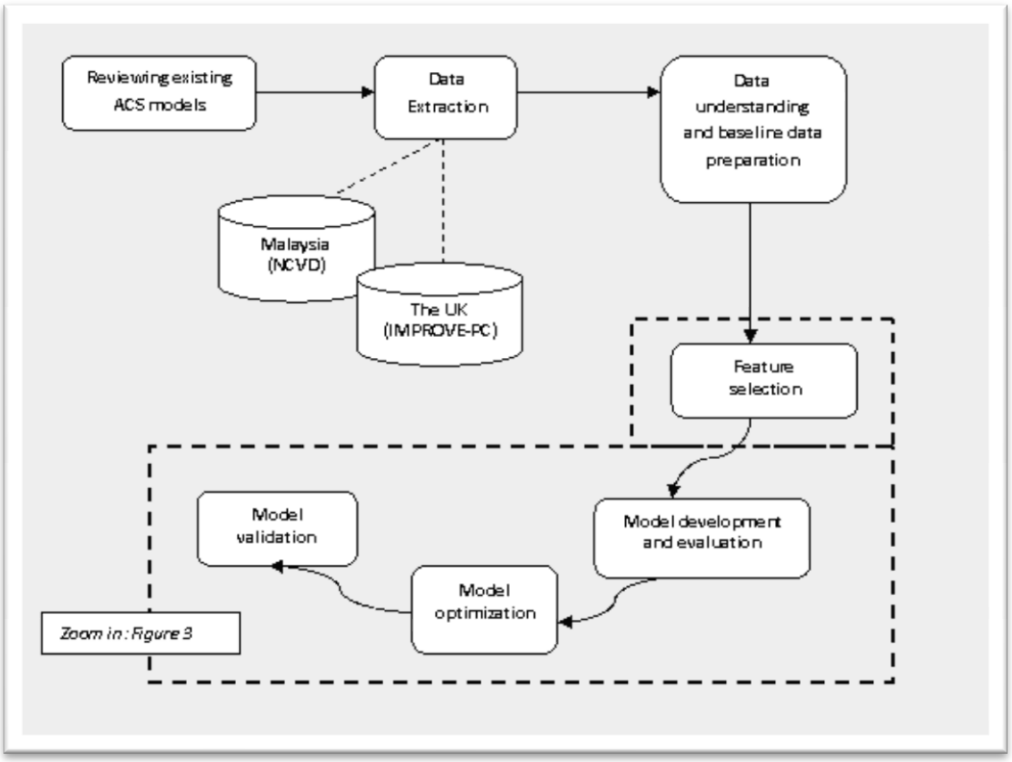


Figure 1: The research methodology

3.1. Evaluation of the Present ACS Prediction Models

The present ACS prediction models are elaborated in the literature. Several aspects, for instance, overall study population, study design, predictors and outcome, prediction methods, and model performances were identified, evaluated, and compared(refer to Section2.1).

3.2. Data Extraction

The data were extracted from two sources, as follows:

1) National Cardiovascular Disease Database (NCVD) - The Malaysian ACS Registry

   A Malaysian national ACS registry has been recording ACS events from 18 hospitals located in Malaysia since 2006.

2) Improving Prevention of Vascular Events in Primary Care (IMPROVE-PC) – The UK ACS dataset

   The dataset is an outcome of the IMPROVE-PC project. The project-linked registry data is from the Myocardial Ischemia National Audit Project (MINAP) with Hospital Episode Statistics (HES) and Primary Care data extracted from nine General Practices (GP) in Leeds for patients diagnosed with Coronary Artery Disease (CAD) from 2000 until 2010. For the purpose of this study, the data information derived from the UK dataset has been limited to MINAP and HES.

The raw datasets retrieved from the two sources were in *.csv* file format.

3.3. Data Understanding and Baseline Data Preparation

The objective of this phase is to understand the overall population in the datasets, potential predictors for model development, and outcomes. Hence, the characteristics of each dataset were defined, and the similarities and differences were evaluated. In addition, data dictionaries for each dataset were used as the main reference to comprehend the overall context of the extracted datasets. The final outcomes of this phase are: 1) the baseline datasets - used to define the population characteristics, and 2) the baseline modelling datasets - for model development.

The summary of baseline datasets and baseline modelling datasets formation are illustrated in Figure 2.

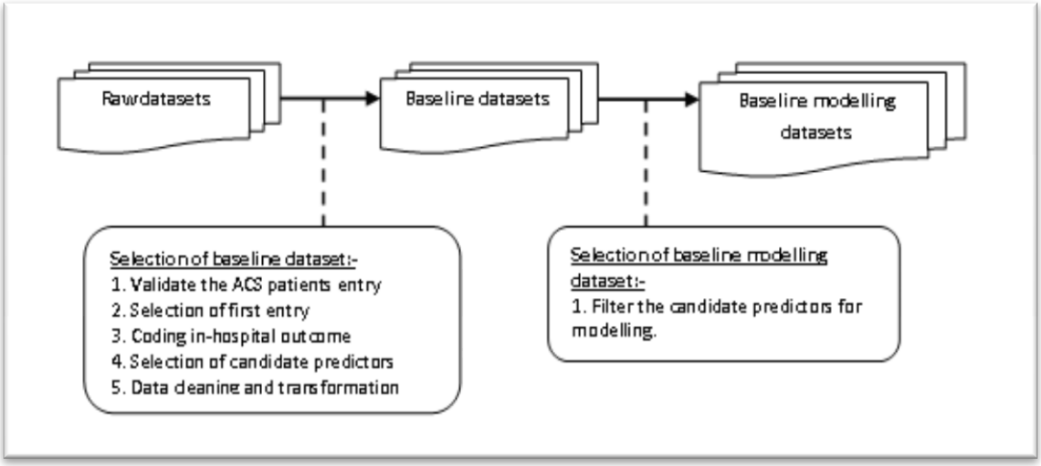


Figure 2: Formation of baseline datasets and baseline modelling datasets

The baseline datasets were extracted from raw datasets. In fact, only ACS entries that met the criteria for each patient entry to the registry at first hospital admission have been considered for this study. The outcome of the model is in-hospital mortality. Thus, relevant in-hospital outcome attributes were identified, coded into binary outcomes, and an appropriate strategy was applied to ascertain missing values. The candidate predictors functioned as attributes with regard to patients' clinical, demographic, and admission information. These attributes were further classified into the following categories:

- Admission
- Demographics
- Status Before Event - Smoking Status, Aspirin Used, Past Medical History, Past Medical Treatment
- Clinical Presentation
- ECG
- Clinical Investigations and Examinations
- Clinical Diagnosis
- Treatment and Interventions
- Medical - Pre Admission, During Admission, Post Admission
- Clinical Outcomes
- Geographical score

Other irrelevant attributes were excluded from the baseline datasets. The datasets were then cleaned up and transformed to ensure the quality of data. As such, graphical charts and descriptive statistics were used to examine the baseline dataset in light of in-hospital mortality. Hence, any potential outlier or noise (miscoding, suspicious data, or just plain error) found within the datasets has been highlighted

After that, using the baseline datasets, baseline modelling datasets were generated. The baseline modelling dataset incorporates only attributes that were used for model development. The model is meant to aid doctors or medical practitioners in making diagnoses for further treatment. The criterion of choosing the attributes is to include only attributes that were captured before making diagnosis or/and any decision regarding the diagnosis. With that, four groups of attributes were excluded, which were:

- Admission
- Clinical Diagnosis
- Treatment and Interventions
- Medical - During Admission, Post Admission
- Clinical Outcomes, except for in hospital mortality

Then, the final baseline modelling dataset was divided into training and testing datasets (i.e., 2/3 training dataset, and 1/3 testing dataset).

During preparation, ‘dirty’ records and ‘dirty’ attributes were deleted. Additionally, some key attributes were imputed, some values in attributes were transformed and several new attributes were created.

3.3.1 Methodology Review

The objective of this task is to review the process of developing the model using WEKA. WEKA is an open-source software developed by the University of Waikato, New Zealand using Java (Witten et al., 2005, Hall et al., 2009). The software, under the GNU General Public License, provides a collections of ML algorithms for DM tasks supported by visualization tools. The software also provides a platform for a developer to develop new DM and ML algorithms. It is globally accepted for data analysis and predictive modelling by many practitioners and research scholars.

A preliminary study was executed using the Malaysia dataset with a smaller number of attributes to compare three ML algorithms in WEKA. The preliminary study was presented in the 33rdd SGAI International Conference on Artificial Intelligent(Jaafar et al., 2013). Also, the objectives of the task were to identify 'unsuitable' ML algorithms for the datasets derived from WEKA involving 29 ML algorithms, explore the effect of missing values in developing a model, and evaluate the varied proportions of the random sampling method.

3.4. Feature Selection

Each Malaysian and UK baseline dataset exceeded 50 attributes in size. The study employed a feature selection method using ML as a technique to simplify the model, yet retain its good predictive power. Different sets of predictors were also grouped to achieve Objective 2 of the study, i.e., to investigates the potency of the current set of predictors in developing ACS prediction using ML techniques, and to investigate the strength of predictors from different clinical categories in producing good predicting models. As such, various sets of predictors were established. These subsets of predictors were employed as input datasets to build the models based on varied learning algorithms.

3.5. Model Development and Evaluation

The complete model development in WEKA embeds the process of preparing the input files, inclusive of feature selection, placement of the input dataset for training, and training the dataset by using a specified classification algorithm (development of a model). Next, the model was tested for its validity with the testing datasets. In this study, the AUC was employed to measure the discrimination performance of the models. Figure 3 portrays the feature selection process, model development, model evaluation, and model validation implemented in this study.

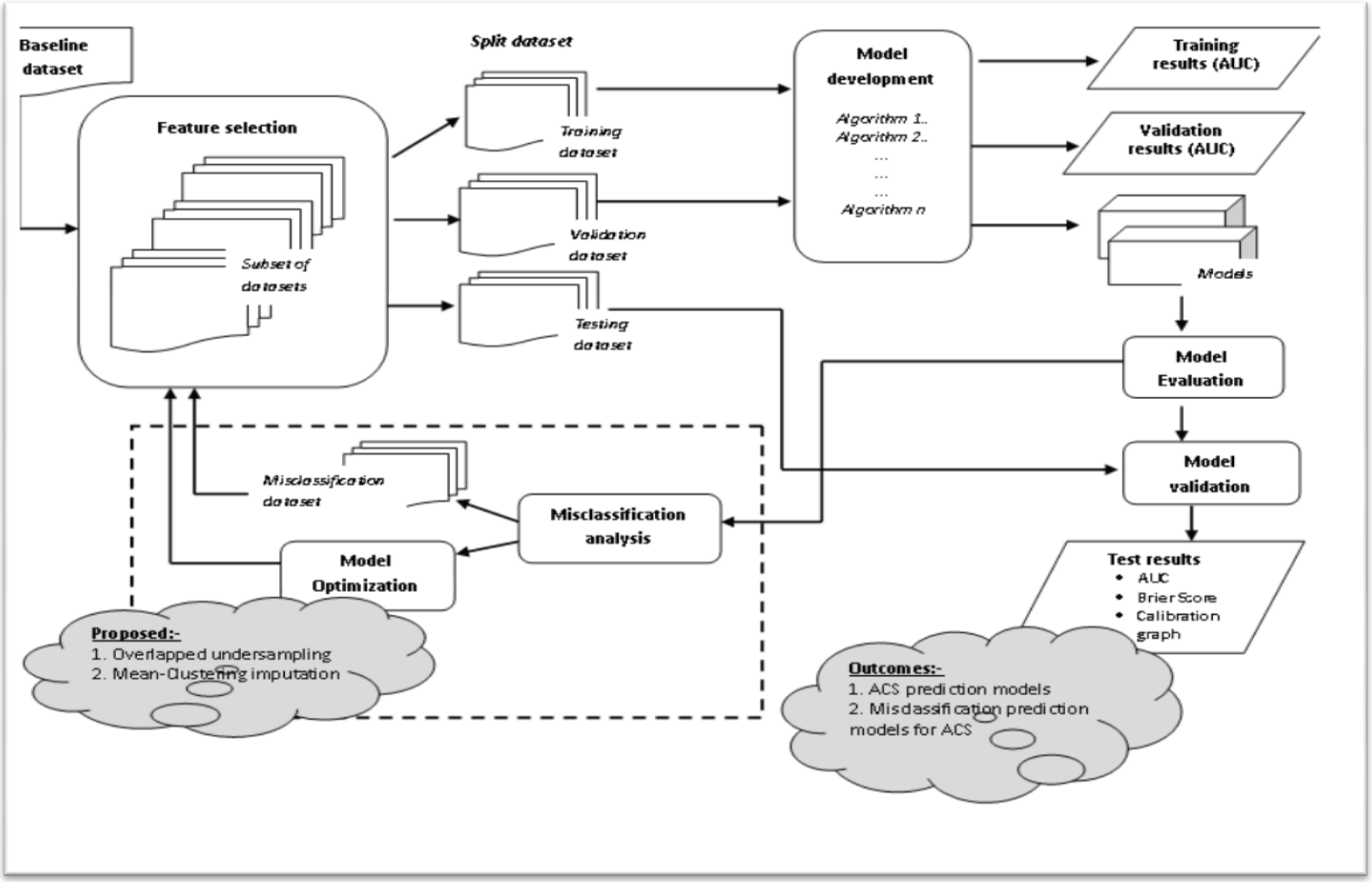


Figure 3: Detailed methodology of model development, model evaluation, and model validation

As depicted in Figure 3, from the baseline modelling dataset, several subsets were extracted as a result of the feature selection task. These subsets functioned as the input datasets to develop the ACS prediction models on various classification algorithms. These models were then validated by using the validation dataset. Next, the AUCs of the models were assessed and compared to determine the best classification models. In fact, some of these models were applied to analyse the reasons for misclassification. Hence, based on the misclassification analysis, the models were further optimized by employing some methods to handle imbalanced datasets and missing values. Furthermore, the varied methods employed to handle imbalanced datasets and missing values were compared with the methods proposed in this study, i.e. the *overlapped undersampling* method and *mean-clustering-imputation* method. In addition, the prediction models that estimated the misclassification instances were also built as a result of the misclassification analysis task. Later, the best classification models were tested on both internal and external datasets (if applicable). Lastly, AUC, Brier score (BS), and calibration plots were compared to identify the best classification models.

3.5.1 Classification Algorithms

One of the objectives of this study is to establish the best ML algorithms for developing ACS prediction models. Subsequently, this study also identified 'unsuitable' algorithms for the dataset, as set out in Objective 1 of the study. As such, 29 ML algorithms with a default parameter setting available in WEKA were evaluated. These classification algorithms originated from distinct basic learning concepts, namely, Naive Bayes, Linear/Non-Linear, SVM, Neural Networks, Instance-based Rules, and Tree models. These are the basic concepts of learning that ML algorithms are developed from.

Furthermore, Gibert et al.(2010) asserted that to select potential ML algorithms for classification modelling, the primary task of the modelling and the dataset structure appear to be the two parameters that must be factored in. In fact, these 29 selected algorithms suit with the study objectives and attributes of the datasets. Moreover, these algorithms support the classification task (prediction capabilities with some algorithms supporting

description capabilities, in which the patterns could be understood by humans) with a dichotomous outcome, as well as both continuous and categorical attributes.

The 29 algorithms as depicted in Section3.3.1were first evaluated. In addition, the 'unsuitable' classification algorithms for the ACS datasets generated in this study were eliminated, while the rest were used for model development. Table 2 presents the evaluated modelling algorithms.

Table 2: List of evaluated classification algorithms

Basic Algorithms	WEKA Modelling Algorithms/Classifiers
Naïve Bayes Learning	Bayes Net (BN) Naïve Bayes (NB)
Linear Models/Non-Linear	Logistic (LG)
SVM	SMO(SVM)
Neural Networks	MultiLayerPerceptron (MLP) VotedPerceptron (VP)
Instance-Based Learning	K-Nearest Neighbour (KNN) Locally Weighted Naive Bayes (LWL)
Rules	Conjunctive Rules (CR) Decision Table (DT) Decision Tables and Naive Bayes (DTNB) Repeated Incremental Pruning to Produce Error Reduction (Jrip) OneRule (OneR) Projective Adaptive Resonance Theory (PART) RIpple-DOwn Rule (Ridor) ZeroR (ZR)
Tree-Based	Alternating Decision Tree (ADT) Bloom-Filter Tree (BFT) Decision Stump (DS) Functional Tree (FT) C4.5 decision tree (J48) Grafted C4.5 decision tree (J48Graft) Logistic Alternating Decision Tree (LT) Logistic Model Trees (LMT) Nave Bayes Tree (NBT) Random Forest (RF) Random Tree (RT) REPTree (REPT) SimpleCart (SC)

The evaluated ML classification algorithms are grouped into seven basic learners as illustrated in Table 2. Algorithms under the NB learner uses the classical statistical theory i.e. Bayes theorem(John and Langley, 1995) as the basis of the algorithm. The LG algorithm in WEKA uses regression technique with a ridge estimator(Witten et al., 2005). On the other hand, SVM algorithm uses the maximum-margin hyper-plane to determine the best

separation for the classes(Vapnik, 1998). Neural Network is a learner which uses the basis of human brain interactions in processing and understanding relationships. As for instance based learning, a distance function is used to determine the shortest distance between the training samples and test samples(Witten et al., 2005). While rule and decision tree learners are based on divide-and-conquer approach which normally work on top-down manner. At each stage, the best identified attribute is split into classes, and recursively process the sub problems resulted from the split. Unlike decision tree, rule based learner comes with a rule in selecting the instances at each stage. Thus, the rule based learner will lead to a set of rules rather than a decision tree(Witten et al., 2005). Different rules, different splitting methods and different pruning strategies (to reduce number of nodes in a tree) differentiate the algorithms under rule and decision tree learners.

3.5.2 Evaluation Methods

3.5.2.1 Hold Out Method - Random Sampling

Due to the massive size of the dataset to be learnt by the classifiers, a hold-out strategy was employed to evaluate performance metrics (Witten et al., 2005). The baseline datasets were randomly divided into two sets: 1) a training set - to construct and evaluate the model, and 2) a test set - to estimate the final performance of the selected model. In the methodology review, various proportions of both training and validation sets were examined for the29 ML algorithms, primarily to identify the best range of hold out method for the datasets.

3.5.2.2 Discrimination

In the clinical prediction models, two primary aspects are considered for measuring model performance, namely, discrimination and calibration (Steyerberg, 2009). Discrimination refers to the ability portrayed by a predictive model in distinguishing outcomes. A perfect discrimination refers to the ability of a model to perfectly place the tested elements into their true classes. One of the widely used performance measures used to evaluate discrimination within the context of dichotomous classification is the receiver operating characteristics (ROC) curve. The ROC curve presents the relative trades-off between true positive (sensitivity) and false positive (1-specificity)

(Pepe, 2003). Apart from a medical clinical decision and diagnostic tests, the ML community recommends discrimination as a performance measure with which to compare prediction models (Bradley, 1997, Fawcett, 2006, Kumar and Indrayan, 2011). The summary measure of ROC is an area under the ROC curve is AUC, which is also known as a c-statistic. Moreover, AUC has been widely used in medical journals to assess predictive performance. Even though some issues have been raised in applying AUC, or the c-statistic(Lobo et al., 2008, Cook, 2007), the application of AUC-ROC has been utilized by various disciplines, including in recent developments in ACS mortality prediction models (Huynh et al., 2013, Kurz et al., 2009).

Therefore, the AUC was used in this study to access the discrimination capability among the models. Besides, using AUC, comparison in terms of predictive performance between the developed models and the existing ACS models is relevant since AUC/c-statistics has been the way of measuring the predictive performance of existing ACS models.

3.6. Misclassification Analysis

The objective of misclassification analysis is to determine the causes of misclassified instances resulted from model development. Hence, this analysis focused on examining misclassified instances in minority classes, overlapping classes, outliers, and missing values. Based on the results obtained from the misclassification analysis, a prediction model that estimated misclassified instances for ACS was developed.

3.7. Model Optimization

Model optimization refers to applying several strategies to increase model performance. The two methods proposed in this study to improve the model performance were: 1) the *overlapped-undersampling* method - to address issues related to imbalanced datasets, and 2) the *mean-clustering-imputation* method - to handle missing values.

3.8. Model Validation

Model validation validates the best models for the datasets. The best models for the datasets go through internal and external validation processes. The calibration measure of the best models in this study was calculated by using the Brier Score (BS). Additionally, calibration plots were applied to illustrate the calibration of these models.

3.8.1 Internal and External Validation

Internal validation incorporates testing a model with similar underlying populations, whereas external validation tests a model on other populations. As for this study, the testing dataset reserved earlier in the data understanding and data preparation processes has been employed to internally validate the models. In order to validate a model, the testing set must display similar features to the derived model. For instance, if the model were built with four features: 1) Age, 2) Heart rate, 3) SBP, and 4) Height, those features must also exist in the testing set. Therefore, the predictors of the best models were decreased to compromise with the external testing dataset. As for external validation, models built based on the Malaysian dataset were tested on the UK dataset, while models developed on the UK dataset were tested on the Malaysian dataset. As a result, a generic ACS prediction model for both the Malaysian and UK datasets had been identified after executing external validation.

3.8.2 Brier Score (BS)

Another essential measurement for prediction models is calibration (Steyerberg, 2008, Van Calster et al., 2015). A well-calibrated model is vital for later use in risk adjustment. In addition, calibration refers to the assessment of how well a model could predict, in comparison to the actual events. In this study, the best models were calibrated by using BS. BS, which is a method proposed by Glenn W. Brier in the 1950s,denotes a scoring rule based on a simple mean squared error of the predicted value (in comparison to the actual outcome).

The formula is presented as below:

$$BS = \frac{1}{N}\sum_{t=1}^{n}(f_t - o_t)$$

Where,

N is the number of samples

f is the predicted probability

o is the outcome (1 if the event occurred, 0 if did not occur)

∑is the summation of the values

A BS value close to zero means good prediction, whereas a score towards 1 indicates otherwise.

In the study, the predicted probability of each instance was obtained from one of the output options provided by WEKA. The sample output from WEKA is depicted in Figure 4. The predicted probability is represented in the "probability distribution" column of Figure 4 which represents the predicted probability of negative and positive cases of each instance.

```
User supplied test set
Relation:     4.2.NCVD_Testing_GRACE_NoMissing
Instances:     unknown (yet). Reading incrementally
Attributes:   7

=== Predictions on test set ===

inst#,     actual, predicted, error, probability distribution
     1 1:Discharg 1:Discharg          *0.973  0.027 (1963)
     2 1:Discharg 1:Discharg          *0.984  0.016 (21180)
     3 1:Discharg 1:Discharg          *0.969  0.031 (15095)
     4 1:Discharg 1:Discharg          *0.973  0.027 (3615)
     5 1:Discharg 1:Discharg          *0.931  0.069 (22715)
     6 1:Discharg 1:Discharg          *0.992  0.008 (17102)
     7     2:Died 1:Discharg      +   *0.969  0.031 (2869)
     8 1:Discharg 1:Discharg          *0.984  0.016 (5956)
     9 1:Discharg 1:Discharg          *0.973  0.027 (1984)
    10 1:Discharg 1:Discharg          *0.958  0.042 (21587)
    11 1:Discharg 1:Discharg          *0.944  0.056 (9919)
```

Figure 4:A snippet of sample output produced by WEKA

3.8.3 Calibration Plots

A calibration graph was generated for each model to visually illustrate models prediction versus the actual outcome. The graph was plotted with the predicted mean values on the x-axis, while the values in the y-axis represented the recorded values.

Due to the binary outcome, using the actual outcomes of 0 or1 fails to provide meaningful variances. Hence, a method of binning, as recommended by John Tukey, has been adopted in this study. The binning method divides the mean values of predicted probabilities in a number of bins with each bin consisting of similar prediction probability values. In this study, the mean values were divided using ten quantiles. As such, the predicted mean value for each of the ten bins was compared with that of the actual outcome. Additionally, a model is said to be well-calibrated if the predicted and the actual mean values for each bin are close in value. Therefore, a well-calibrated model should show the points on the graph lying close to the 45-degreediagonal line. Moreover, plotting these values on a graph offers a better view of the ability of the predicted values to calibrate the observed values.

Chapter 4: Data Understanding, Data Preparation, and Methodology Review

This chapter describes the characteristics of the two datasets employed in this study, derived from Malaysia and the UK. In addition, the chapter elaborates on the processes of preparing the baseline and baseline modelling datasets. Furthermore, the chapter explains the process of reviewing model development using WEKA, as well as several strategies applied in model development.

4.1. Overview of ACS Datasets

4.1.1 Malaysian Dataset - National Cardiovascular Disease Database (NCVD)

The Malaysian ACS registry, which is hosted by NCVD, is supported by the Malaysian Ministry of Health (MOH). This database is central and 'live,' as it integrates all CVD databases in Malaysia to strategically manage CVD treatment and improve overall cardiac services in Malaysia. Given this, the registry is comprised of information pertaining to patients diagnosed with ACS, including STEMI, NSTEMI, and UA, aged 18 years old and above, and admitted to one of 18 participating sites throughout Malaysia.

As a standard procedure in Malaysia, a patient record is created upon admission to the hospital. Additionally, records in the ACS registry are captured from patients' records at the hospitals. In fact, it was ascertained that these records satisfied the ACS enrolment criteria before being transferred to the ACS registry. Moreover, all the records were first validated and cleaned before being transferred into the registry. The data stored in the registry were designed by the content experts in the discipline, led by a team from the Cardiology Department of the MOH, universities, National Heart Institute, and Department of Medicine at the Kuala Lumpur Hospital. The information was designed based on international registries and guidelines issued by the Australian National Data Elements for ACS, the European Cardiology Audit and Registration Data Standards (CARDS), as well as the American College of Cardiology Clinical Data Standards(Chin et

al., 2008). Follow-up data of 30 days and 12 months upon initial registration were also captured in the ACS registry.

The information in the registry includes admission details, demographics, past medical history, clinical and procedure information, as well as pharmacotherapy. Since the data is derived from the registries of 18 hospitals that cover all 14 states in Malaysia, both ACS events and trends in the registry were assumed to reflect ACS events and trends throughout Malaysia, mirroring the varied races and ethnicities in Malaysia(Ahmad et al., 2011).

4.1.2 The Leeds, UK Dataset - Improving Prevention of Vascular Events in Primary Care (IMPROVE-PC)

The IMPROVE-PC dataset refers to the outcome from a small part (Cardiovascular Healthcare Information Linkage Study) of the overall IMPROVE-PC project. The main aim of the IMPROVE-PC project has been to promote healthy lifestyles by changing the behaviour among those with a high risk of being diagnosed with CVD in the Leeds area. Hence, to attain the aim, it is crucial to find patients at high risk, which can be done by going through patient records in GP and/or hospitals. As such, data recorded in both health care systems must be of good quality, reliable, complete, and consistent. Thus, the Cardiovascular Healthcare Information Linkage Study is a project that links both primary and secondary care data in Leeds in order to evaluate the quality of recorded data in both health care systems(CLAHRC for Leeds).

In fact, the project linked three databases, which were the: 1) GP - SystemOne primary care, 2) MINAP, and 3) HES. MINAP denotes the registry for hospital admission records for all ACS patients in England and Wales, whereas HES is composed of details of hospital episodes in NHS hospitals within England and all other hospitals that offer services to NHS patients. The selected sample for the linkage studies included patients diagnosed with CVD, who had Leeds postcode, were registered under a selected GP using SystemOne, and were registered as an inpatient and outpatient at the hospitals(House et al., 2011).

4.2. Data Extraction

Data extraction refers to the process of extracting datasets from the registries.

4.2.1 The Malaysian Dataset

In order to employ the Malaysian data for this study, initially, the data had to be requested from and approved by the NCVD board. As such, the data were approved on 3rd October 2012 and released on 19th October 2012 through a secured network protocol equipped with a password. The secured network protocol was open for extraction for 7 days (19thuntil 25th October 2012), and, on 26th October all the contents were removed. The dataset was securely stored and processed in a private storage (m drive) located at the University of Leeds, which could only be accessed by the main researcher. All the records were anonymized and were saved in a *csv* file. In fact, the details of the NCVD have previously been published(Chin et al., 2008). A total of 13,591patient records from year 2006-2010 were extracted with 215 attributes in each record.

4.2.2 The Leeds, UK Dataset

Similarly, the UK dataset was requested from and approved by Steve Magare, the data manager for the linkage project. In fact, data were requested for HES- and MINAP-linked data for the years 2000 to 2010 for first admissions only. Hence, the dataset only consists of ACS patients derived from the MINAP registry, and additional information on the attributes obtained was from HES. The data were approved and released on 10thOctober 2013 through a secured network protocol, together with secured encrypted password. The data were securely stored and processed in private storage located at the University of Leeds (m drive), which could only be accessed by the main researcher. All records were anonymized and saved as a *csv* file. The total records gathered were 50,588 records with 236 attributes in each record.

4.2.3 Attributes of The Datasets

Each entry of the raw dataset (Malaysian and the UK) is identified by its unique id. The attributes of the datasets are inclusive of clinical, non-clinical, and database-specific data. The clinical information covers data related to admissions, demographics, past medical history, clinical and procedure information, as well as pharmacotherapy, which were further divided into various clinical categories. Meanwhile, non-clinical information refers to data that are unrelated to ACS, whereas database information denotes attributes used for database and meta-data purposes. The attributes are comprised of numerical (discrete and continuous), categorical (ordinal and nominal), date, and text data types. The detailed summary of attributes for the raw datasets are summarized in Appendix A, withA.1.1 *Summary of Attributes* describing the Malaysian dataset, and A.2.1 *Summary of Attributes* describing the UK dataset.

4.3. Data Preparation

4.3.1 Baseline Dataset Preparation

A baseline dataset was used to comprehend the populations in the study, as well as to statistically summarize the characteristics of the study populations. Baseline datasets is a subset of the raw dataset. Moreover, these baseline datasets were the outcome of cases filtered in accordance to the study scope, formation of the outcome attribute, and selected candidate predictors, as well as the overall cleaned up and transformed datasets.

4.3.1.1 Study Population

The sources of both the Malaysian and UK datasets were registries, which were mainly derived from hospital records. Before the hospital records were transferred into registries, specific data validation and cleaning procedures were executed (Gale et al., 2008a, Chin et al., 2008). Hence, an assumption was made that all records in the dataset were related to those diagnosed with ACS based on the specification of the registries. Nonetheless, the data were still validated to ascertain that only patients 18 years old and above and admitted between 2006 and 2010 for Malaysian

dataset, and 2000 and 2010 for the UK dataset, were included. All records that failed to meet the criteria were excluded from the dataset.

4.3.1.2 Selection of First Entry

In addition, it is important to note that the study only considered the first hospital admission for each patient. All subsequent entries were considered duplicates, and were thus excluded from the dataset.

As for the Malaysian dataset, multiple admission dates of a patient were used to identify duplicate patients. As such, 669 patients were detected with multiple entries. Thus, the first admission date was considered as the first entry, while the remaining entries were considered as subsequent entries, and hence excluded from the dataset. However, 29 patients with multiple entries had a similar admission date. For such cases, the entries were first evaluated to determine the available information that led to the identification of first entry. If this did not work, the strategy was to look for an entry with more valuable information, for example, fewer missing values and less noisy data. If the entries shared similar valuable information, the one with the lowest notification id was selected as we assumed that the notification id was generated in an incremental manner. There were also 11 cases in which admission was detected at two different hospitals either on the same date or on two consecutive days. These particular cases were due to immediate transfer of patients to another hospital. Hence, the entry for the second hospital was selected, primarily because the patients stayed longer at the second hospital, when compared to the initial hospital (discharged on the same date or the next after being admitted), thus suggesting more data collection during the stay. Moreover, it is also notable that all these cases referred to patients who were transferred to the National Heart Institute or a specialized hospital for heart problems. Thus, it was assumed that more reliable and thorough data were collected from these hospitals. On top of that, records for four patients had similar admission dates, but the information between the records were totally sparse, thus leading to dead ends. Hence, these records were excluded from the dataset. Finally, there were two other special cases which, in each case, involved the same patient, but had varied admission dates with

overlapping durations of stay at two hospitals. These records were also deleted from the dataset.

As for the UK dataset, a request was made to only extract the first admission entry for each patient. Therefore, the records retrieved were assumed to contain no duplicates. Besides, no attribute in the dataset could function as an indicator of duplicate patients.

4.3.1.3 Preparation of In-Hospital Outcome

The dichotomous, categorical outcome for the prediction model is either dead or not dead. In addition, in supervised learning or classification modelling, it is required that there be no missing values for the outcome, as the correlations between predictors and outcome cannot be analysed, but such relationships are indeed of key interest. Moreover, in studies pertaining to prediction, cases with missing outcomes are generally discarded.

In terms of the Malaysian dataset, the attribute that reflected the in-hospital outcome was *ptoutcome*. The attribute is of type categorical, with two values: 1) Died – indicating that the patient died during his/her stay at the hospital, and 2) Discharged – indicating that the patient lived to leave the hospital. Nonetheless, some 353 records were found to have missing outcomes in the dataset. Each record in the Malaysian dataset has dead or alive information attached, that was recorded by the National Registration Department of Malaysia (NRDM). NRDM is a department within the Malaysian Ministry of Home Affairs that records and manages each important event in the life of an individual in Malaysia, such as birth, death, marriage, divorce, and citizenship status. Therefore, the dead or alive information from NRDM was applied to impute the missing in-hospital outcome for the dataset. Nevertheless, only records with 'Not Died' (alive) status recorded in NRDM were considered for imputation. This is because those that were listed as deceased could have died for reasons unrelated to the ACS event recorded in the registry. After deliberate consideration, 309 records with missing outcomes were imputed with a 'Discharged' value, while all other records were deleted from the dataset.

When considering the UK dataset, the in-hospital outcome attribute was referred to as*X404.Death.in.Hospital*, which was derived from MINAP. The attribute is of type categorical with seven different values, namely, 1) No - indicating that the patient had not died, 2) From MI – indicating that the patient died due to MI, 3) From complication of treatment – indicating that the patient died due to complications from treatment, 4) Other cardiac cause – indicating that the patient died due to other cardiac issues, 5) Other non-cardiac cause – indicating that the patient died due to a non-cardiac cause, 6) Unknown – status is unknown, and 7) NA – unavailable information. The missing values in the attribute are indicated with 'Unknown', 'NA,' or a blank. No attributes in the datasets could possibly suggest a reliable imputed value for the outcome attribute. As such, a total of 45,328 records were considered as missing and were deleted from the dataset.

4.3.1.4 Selection of Candidate Predictors

Candidate predictors were selected by identifying relevant attributes for the research. All attributes pertaining to ACS and clinical elements were considered as potential predictors. These attributes, even excluded from model development ,were important for dataset characteristics and generalization analyses. After careful analysis, six categories of attributes were considered to be irrelevant to the research objectives outlined, and were thus discarded.

1. Duplicate attributes

   Some attributes were notably duplicates. For example, the attributes *contactinstitutionname* and *sdpid,* in which *contactinstitutionname* reflected the hospital/centre, while *sdpid* denoted the code for each hospital/centre. They were a one-to-one match.

   Refer to Appendix A for lists of the dataset attributes that were duplicates. In particular, seeA.1.2 *List of Duplicate Attributes* and *A.2.2 List of Duplicate Attributes* for the Malaysian and UK datasets, respectively.

2. Database-related attributes

   These attributes were irrelevant to model development or for any analysis purposes.

   Refer to Appendix A for lists of irrelevant attributes. In particular, seeA.1.3 *List of Database Attributes* and

   A.2.3 *List of Database* Attributes for the Malaysian and UK datasets, respectively.

3. Unknown attributes

   These attributes did not reflect clinical elements and did not provide any information. In addition, these attributes were not described in the data dictionary or data definitions.

   Refer to Appendix A for further information. In particular, see

   A.1.4 *List of Unknown Attributes* and

   A.2.5 *List of Unknown Attributes* for the Malaysian and UK datasets, respectively.

4. Irrelevant attributes

   These attributes were irrelevant for modelling or generalization in this study. The specific reasons for excluding these attributes are described in Appendix A.

   Refer to A.1.5 *List of Irrelevant Attributes and*

   A.2.5. *List of Irrelevant Attributes* for the Malaysian and UK datasets, respectively

5. Non-standardized data collection attributes

This particular category of attributes only existed in the Malaysian dataset. The attributes referred to information in which the values were neither standardized nor reliable for this study. For instance, information concerning Peak Troponin was captured in various manners, depending on the equipment used in the hospitals/centres. Additionally, different equipment led to different values.

Refer to Appendix A:

A.1.6 *List of Non-standardized Data Collection* Attributes.

6. Dependent attribute for missing values

This scenario only existed in the Malaysian dataset. The missing values were represented in varied ways, such as specifying the missing value with "Missing," "Not Available," "Unknown," or simply a 'blank.' However, missing values were also represented by specifying values in another dependent attribute. In actual fact this dependent attribute is duplicate of the actual attribute. For example, attribute *heightna* refers to an attribute that records the missing values of patient's height. However, if the height of a patient is not captured or found missing, the value was represented as 'blank' in the attribute *height*. These 'blank' values in *height* indirectly represents the information *heightna* was supposed to capture.

Refer to Appendix A:A.1.7 *List of Dependant Missing Attributes*..

4.3.1.5 Data Cleaning and Transformation

New attributes were created to simplify the existing information. Refer to Appendix A: A.1.8 *List of New Attributes* and A.2.7 *List of New Attributes* for the Malaysian and UK datasets, respectively.

4.3.2 Baseline Modelling Dataset Preparation

The baseline modelling dataset refers to baseline datasets meant for modelling. This incorporates a process of reducing the candidate predictors in a baseline dataset to suit the objectives outlined for model development. In this case, the model is targeted to help doctors or medical practitioners in making a diagnosis for further treatment. Hence, this study only considered

attributes that were captured before making any decision on diagnosis and decision about the diagnosis. As a result, attributes from the following categories were discarded.

- Admission
- Clinical Diagnosis
- Treatment and Interventions
- Medical - Post Admission
- Clinical Outcomes - Advice, Rehab, Therapy(Only applicable in the UK dataset)

4.4. Results

4.4.1 Baseline Datasets

After filtering, selecting, and cleaning up the cases and attributes, a total of 12,710 records with 75 attributes were left for the Malaysian baseline dataset (down from 13,591 records with 215 attributes),and 5,127 records with 65 attributes were left for the UK baseline dataset (from an original 50,588 records with 236 attributes). The final baseline dataset resulted from the deletion of approximately 90% of the raw dataset. The large number of excluded records from UK dataset was due to quality issues, as mentioned earlier.

Although a large number of records were deleted from the raw dataset, the sample size of the UK dataset (n=5,127)is still considered appropriate for prediction modelling. In terms of ML classification modelling, Mukherjee et al. (2003) identified that, in the treatment outcome problem, the minimum size for a training sample for a classification problem is more than 50 observations (Mukherjee et al., 2003). For validating a ML classification model, Beleites et al.(2013) suggested that, in order to validate a model from a small sample (i.e., 25 samples per outcome), a minimum of 75-100 observations are needed. Moreover, Hu et al.(2016) demonstrated how they had achieved a good performance model with a small sample size. The study evaluated the effect of different sample sizes on the AUC by modelling with different sample sizes. The result illustrated that at 20% (586/2930) of

the sample size, the performance of the models were found to be relatively stable. Thus, the sample size of the UK dataset is sufficient to produce a reliable model. Two thirds of the baseline datasets were reserved for model derivation, while the remaining one third of the datasets were reserved for model validation.

The variation in attributes between the Malaysian and UK datasets was mainly due to the slightly varying levels of information captured from the datasets. Otherwise, most of the attributes reflected standard information for ACS. As such, 31 common attributes were generated from both the Malaysian and UK datasets. Refer to Appendix A:A.3*The Mapping of Malaysian and The UK Datasets* for a list of common attributes.

4.4.1.1 Data Quality Issues

Generally, as mentioned, both datasets had issues related to duplicates and unknown attributes, as well as missing data. The duplications of attributes in the UK datasets were mainly due to the varied sources of data, such as MINAP and HES. Both sources of data possessed their own attributes, which stored similar information. Meanwhile, unknown attributes refer to attributes that are not specifically defined in the data dictionary. In fact, this case could be suggested to the contributors of datasets so as to improve their data dictionary specification, and, probably, the overall database design. Furthermore, many attributes had missing values for both datasets. Moreover, after applying missing value analysis to the baseline datasets, no complete case was detected in the UK dataset, while only 317 complete cases were found in the Malaysian dataset. The UK dataset might not have any complete case due to the lower number of cases in the dataset in comparison to that in the Malaysian dataset.

The study has identified that, within the Malaysian dataset, several attributes, such as tropinin and creatine kinase MB (CK-MB), were not captured via standard metrics, as different hospitals or centres employed varying forms of metrics. Hence, although these attributes were vital as candidate predictors, they were discarded mainly for being described by non-standardized values. Furthermore, data standardization is a key criterion towards attaining the maximum benefits of a registry (Workman and

A, 2013). Therefore, the findings from this study will be brought to the attention of the NCVD in order to improve the overall data collection strategy. On the other hand, for the UK dataset, some attributes only consisted of a single value, which was mainly associated with "Not Applicable." This indicates that the attributes might not be applicable for the population in the study. In addition, outliers, such as in the attributes SBP, height, weight, and cholesterol reading, were also noted in the UK dataset, were eliminated before initiating model development.

4.4.1.2 Study Population Characteristics

Table 3: Baseline characteristics of both the Malaysian and UK datasets

Characteristics	Malaysian dataset		UK dataset	
	2006 - 2010		2003 - 2010	
	Derivation (n= 9533)	Validation (n= 3177)	Derivation (n=3845)	Validation (n=1282)
Age (years)	59.0 (12.1) [0%]	58.7 (11.9) [0%]	68.8 (13.4) [0%]	68.9 (13.2%) [0%]
Male	7225 (75.8%) [0%]	2439 (76.8%) [0%]	2464 (64.1%) [0%]	817 (63.7%) [0%]
SBP	139.1 (28.7) [1.7%]	139 (28.7) [1.9%]	147.8 (242.8) [23.1%]	143.5(29.6) [23.8%]
Height	161.7 (8.3) [45%]	162.2 (8.2) [45.4%]	166.1 (65) [70.9%]	163.1 (26.1) [73.3%]
Weight	67.6 (14.1) [38.1%]	68.2 (14.1) [38.6%]	78.3 (18.2) [60.6%]	79.4 (42.6) [62.8%]
Heart rate (beats/mins)	83.6 (21.3) [1.7%]	83.7 (21.3) [1.7%]	83.7 (34.8) [23.1%]	85.2 (27.4) [23.7%]
Total Cholesterol	5.31 (1.3) [28%]	5.3 (1.4) [27%]	11.8 (140.9) [40%]	6.7 (29.7) [41%]
Killip:> 1	2264 (23.8%) [24.6%]	758 (23.8%) [10.1%]	NA	NA
Previous MI	1569 (16.5%) [20.8%]	524 (16.5%) [21.7%]	2623 (22.1%) [9.7%]	287 (22.4%) [10.9%]
History of heart failure	616 (6.5%) [17.2%]	208 (6.5%) [18.6%]	207 (6.5%) [17.3%]	75 (5.8%) [18.0%]
History of stroke(cerebrovascular)	328 (3.4%) [19.5%]	118 (3.7%) [20.6%]	272 (7.1%) [18.1%]	82 (6.4%) [18.7%]
History of peripheralvascular disease	74 (1.0%) [20.7%]	19 (0.6%) [21.7%]	195 (5.9%) [13.0%]	62 (4.8%) [14.4%]
History of renal failure	586 (7.6%) [19.4%]	185 (5.8%) [21.0%]	159 (5.0%) [18.0%]	62 (4.8%) [18.3%]
Aspirin taken	3056 (32%) [9.7%]	990 (31.2%) [10.6%]	815 (21.2%) [6%]	255 (19.9%)[7.4%]
History of hypertension	5773 (60.6%) [13.8%]	1878 (59.1%) [14.4%]	1566 (40.7%) [10.6%]	513 (40%) [11.9%]
Current smoker	3231 (33.9%) [5%]	1076 (33.9%) [6%]	1009 (26.2%) [12.7%]	336 (26.2%) [14.2%]
History of diabetics	3964 (41.6%) [17.1%]	1318 (41.5%) [17.7%]	567 (14.8%) [8.9%]	227 (17.7%) [9.2%]
BB Given	2269 (27%) [11.9%]	698 (22.0%) [12.5%]	1654(60.5%) [28.9%]	535 (41.7%) [30.5%]
Statin Given	2724 (32.3%) [11.6%]	874 (27.5%) [12.5%]	1993 (72.6%) [28.6%]	658 (51.3%) [30.7%]

Previous PCI	NA	NA	348 (9.1%) [18%]	111 (8.7%) [18.6%]
Previous CABG	NA	NA	259 (6.7%) [16.8%]	78 (6.1%) [17.7%]
ECG				
ST elevation Level 1	1910 (20%) [0%]	615 (19.4%) [0%]	1431 (37.2%) [4.7%]	454 (35.4%) [6%]
ST elevation Level 2	3146 (33.3%) [0%]	1066 (33.6%) [0%]		
Q Wave	NA	NA	NA	NA
ST Depression	2488 (26.1%) [0%]	815 (25.7%)	580 (15.1%) [4.7%]	180 (14%) [6%]
T Wave	2104 (22.1%) [0%]	684 (21.5%)	530 (13.8%) [4.7%]	180 (14%) [6%]
BBB	475 (5%) [0%]	136 (4.3%)	NA	NA
LBBB	NA	NA	171 (4.4%) [4.7%]	60 (4.7%) [6%]
Diagnosis				
STEMI	4651 (48.8%) [0%]	1284 (40.4%) [0%]	1342 (34.9%) [15.1%]	414 (32.3%) [17%]
Non-STEMI	2653 (27.8%) [0%]	695 (21.9%) [0%]	1811 (47.1%) [15.1%]	607 (47.3%) [17%]
UA	2229 (23.4%) [0%]	845 (26.6%) [0%]	116 (3.1%) [15.1%]	42 (3.3%) [17%]
No of stays in the hospital	4.9 (3.6) [16%]	4.7 (3.4) [16%]	7.1 (9) [0%]	7.82 (11.6) [0%]
Treatment				
Fibrinolytic therapy	3430 (88.33%) [0%]	1110 (86.4%) [0%]	65 (1.7%)[46%]	24 (1.9%)[45.5%]
Cardiac catherization	1871 (19.6%) [0%]	609 (19.2%) [0%]	NA	NA
CI	1422 (14.9%) [0%]	467 (14.7%) [0%]	534 (13.8%) [46.0%]	164 (12.8%) [45.5%]
CABG	112 (1.2%) [0%]	39 (1.2%) [0%]	44 (1.1%) [46.0%]	16 (1.2%) [45.5%]
Died - In hospital	681 (7.1%) [0%]	221 (7%) [0%]	184 (4.8%) [0%]	57 (4.4%) [0%]

Values are number (%) or mean (standard deviation) [% of missing values]

As presented in Table 3, Malaysian patients were found to be smaller by weight and height, in addition to being strikingly younger (mean age of 58 years old) when compared to patients from the UK (mean age of 69 years old). Additionally, both populations had a higher proportion of male patients, and the Malaysian dataset recorded about 10% more male patients than found in the UK dataset.

Overall, there were noticeable more patients with a history of MI, hypertension, and diabetes than other types of medical history in both populations. Nonetheless, the Malaysian population had extremely high percentages of patients with hypertension (~61%) and diabetes(~42%). Meanwhile, the prevalence of hypertension (~41%) was the highest among the UK population.

Furthermore, patients taking aspirin were more common in the Malaysian population, while patients taking statins and BBs were more prevalent in the UK population.

Other than that, those diagnosed with STEMI (~48%) were more noticeable in the Malaysian population, as compared to that in the UK (~35%). In fact, the UK population was found to have more prevalence of NSTEMI (47%) cases and a very limited number of UA cases (~3%). The Malaysian population, however, had patients diagnosed with NSTEMI and UA in almost similar percentages (~22-27%).

Out of the 4,651 Malaysian patients diagnosed with STEMI, 2,832 (60.9%) were treated with fibrinolytic therapy, 991 (21.3%) had cardiac catherization, 877 (18.9%) had undergone PCI, and only 24 (0.5%) had CABG. Meanwhile, PCI (28.6%) and thrombolytic therapy (13.4%) emerged as the main procedures used for those diagnosed with STEMI in the UK population.

In addition, the mean duration of hospital stay among UK patients was higher than that of Malaysian patients. Nevertheless, in-hospital mortality among Malaysian ACS patients was higher (7%) in comparison to that in the UK (4%). Furthermore, among the three ACS spectrums, those diagnosed with STEMI and NSTEMI had higher mortality rates in both cohort studies.

4.4.1.3 The Leeds, UK Population Representativeness

As the UK dataset only involved patients from a particular part of the UK, i.e., Leeds, the sample may not represent the UK as a whole. To validate the representativeness of the studied dataset, the dataset was compared to the dataset used in the study by Gale et. al.(2008b), which utilized the MINAP dataset covering all patients in England and Wales from 2003-2005. As described in Section 4.2.2,the Leeds, UK dataset mainly consists of ACS patients from 2003-2010 of MINAP data. Since the datasets are from the MINAP registry, the comparison of these two sets of data is relevant. Table 4 compares the demographic characteristics, medical history, and presenting clinical features used in Gale et. al.'s (2008b) study and this study.

Table 4: Comparison of the Gale et. al. (2008b) MINAP dataset and studied dataset in terms of demographic characteristics, medical history, and presenting clinical features

Characteristics	MINAP (Gale et. al. (2008b)	The Leeds, UK
Demographics		
Age, years (mean (SD))	68.9 (13.79)	68.8% (13.4)
Female	36 198 (36%)	1381 (35.9%)
White	76 111 (76%)	2585 (67.2%)
Asian	3234 (3%)	105 (2.7%)
Medical history		
Myocardial infarction	22 638 (22%)	2623 (22.1%)
Hypertension	42 528 (42%)	1566 (40.7%)
Angina	32 029 (32%)	1017 (26.4%)
Chronic renal failure	3109 (3%)	159 (5.0%)
Cerebrovascular disease	7482 (7%)	272 (7.1%)
Peripheral vascular disease	4319 (4%)	195 (5.9%)
Heart failure	5889 (6%)	207 (6%)
Diabetes	17 125 (17%)	567 (14.8%)
Smoking	25 164 (25%)	1009 (26.2%)
Cardiac enzymes		
Elevated CK or troponin	70 378 (70%)	3103 (80.7%)
ECG changes		
ST-segment elevation	33 723 (33%)	1431 (37.2%)
LBBB	5068 (5%)	171 (4.4%)
ST-segment depression	13 023 (13%)	580 (15.1%)
T-wave changes only	13 020 (13%)	530 (13.8%)
Arrhythmia or conduction abnormality	13 248 (13%)	473 (12.3%)
Normal	8870 (9%)	247 (6.4%)
Aspirin status		
Already taking aspirin before admission	22 363 (22%)	815 (21.2%)

Comparing these two sets of MINAP data, no obvious differences in most of the patients' characteristics were observed. As illustrated in Table 4, the noticeable differences of patients' characteristics are highlighted in grey. Demographically, there were about 10% more white patients in Gale et. al.'s (2008) sample than in the studied dataset. There were also 6% and 3% more patients with history of angina and diabetics, respectively, in Gale et. al.'s (2008) dataset. However, in comparison to Gale et. al.'s (2008) dataset, the studied dataset has about 10% more patients with elevated cardiac enzymes during an acute phase. Having more patients with cardiac enzyme in the studied dataset may possibly explain the higher number of patients with ECG- ST-segment elevation in the studied dataset (>4% higher than in Gale et al.'s (2008) dataset), and the fewer number of patients with normal ECG(>3% more with abnormal results than in Gale et. al.'s (2008) dataset). Thus, from the comparison of these two sets of MINAP datasets, the Leeds, UK dataset is assumed to reflect the UK (Western) population, as a whole.

4.4.2 Baseline Modelling Datasets

Upon filtering candidate predictors, a total of 75 attributes were chosen for the Malaysian baseline modelling dataset, while 65 attributes were chosen for the UK baseline modelling dataset. Each baseline modelling dataset had one id attribute and one outcome attribute, whereas the remaining attributes were considered as candidate predictors. Figure 5 illustrates the distribution of candidate predictors by their clinical categories.

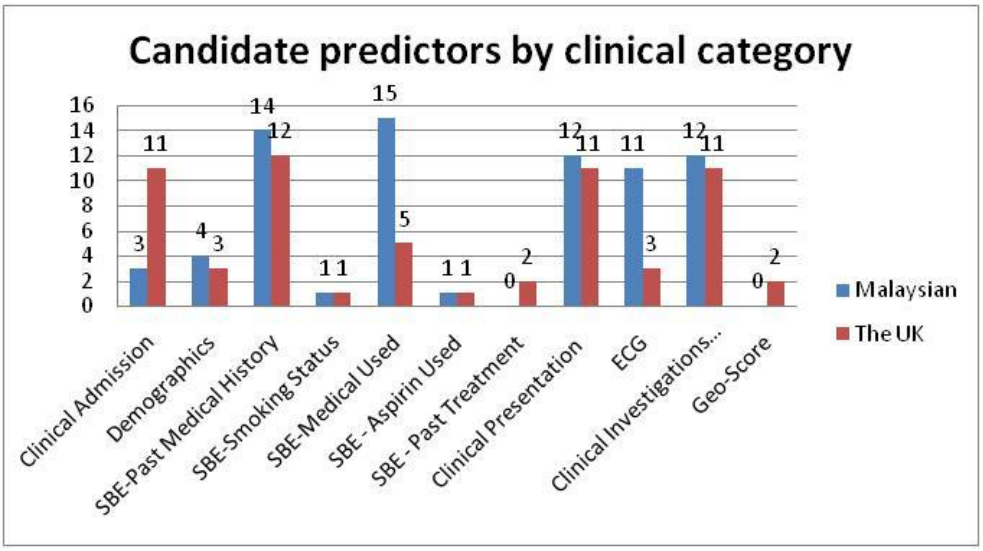


Figure 5: Candidate predictors by clinical categories

SBE- Status before event; Geo-Score- Geographical Score

The distinct number of attributes in the clinical category was observed at clinical admission, status before event-medical used and ECG categories; otherwise they were about the same. An extensive list of medication history for each patient was recorded in the Malaysian dataset, and detailed information for clinical admission was captured in the UK dataset. Nevertheless, when considering the ECG category, for example, the information on the ECG is similar across datasets, and the different number of attributes were due to the distinct way of storing the information in the registries. In addition, no geographical score was stored in the Malaysian dataset. Approximately 77% of the attributes in the baseline modelling datasets were categorical variables, while the rest were numerical.

The missing values were highlighted as a quality issue among the datasets. Figure 6 presents the distribution of missing values, while Figures 7 and 8portray the missing values by their clinical categories.

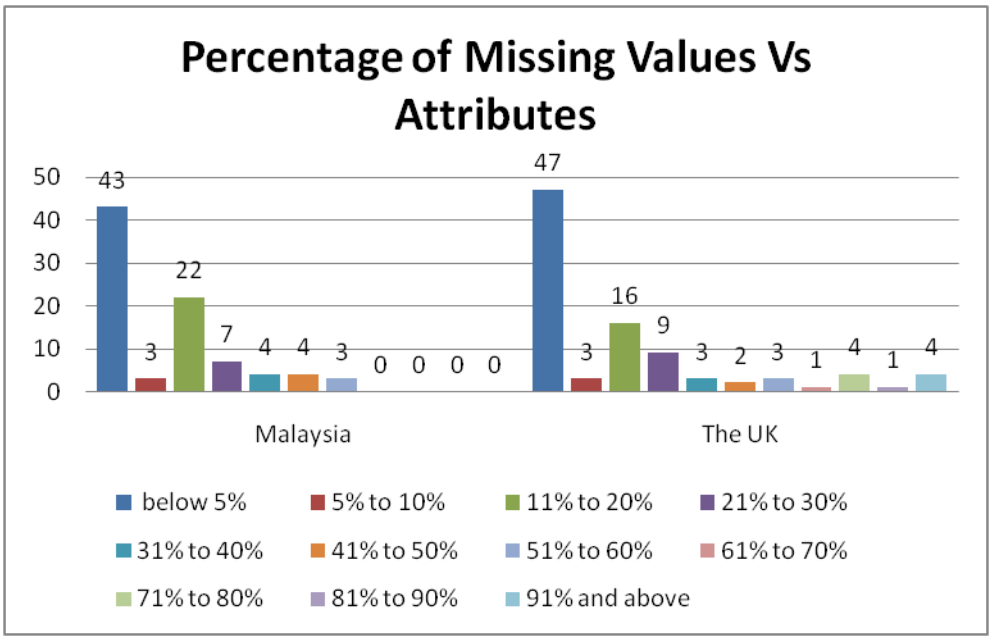


Figure 6: The number of attributes with the given percentage of missing values

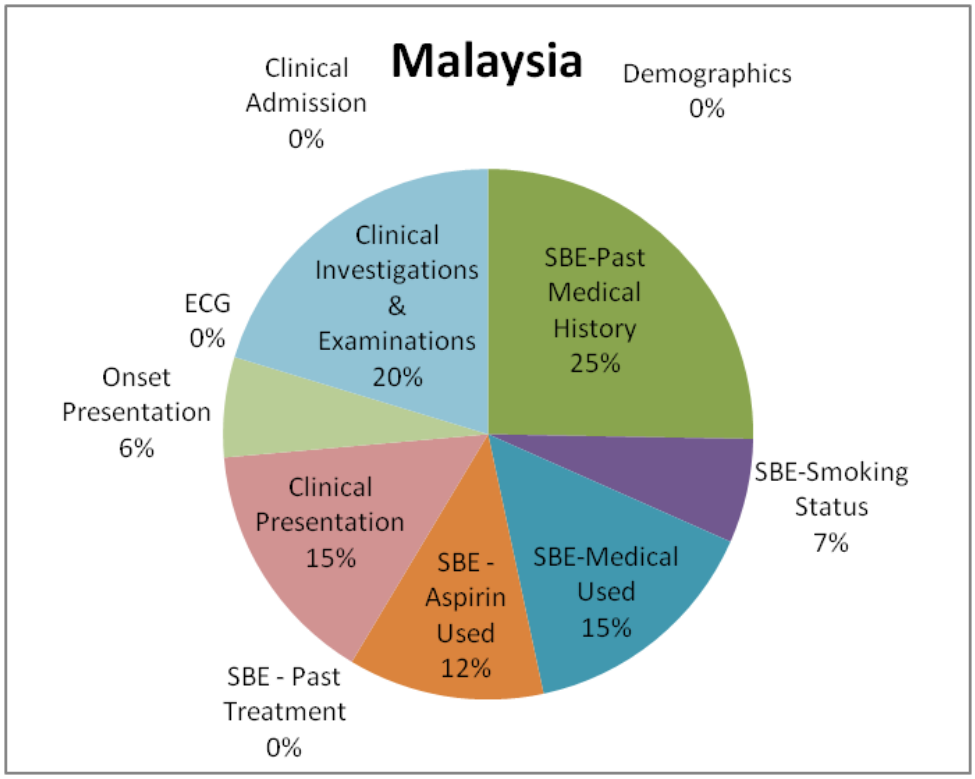


Figure 7: Mean percentage of missing values by clinical category - The Malaysian dataset

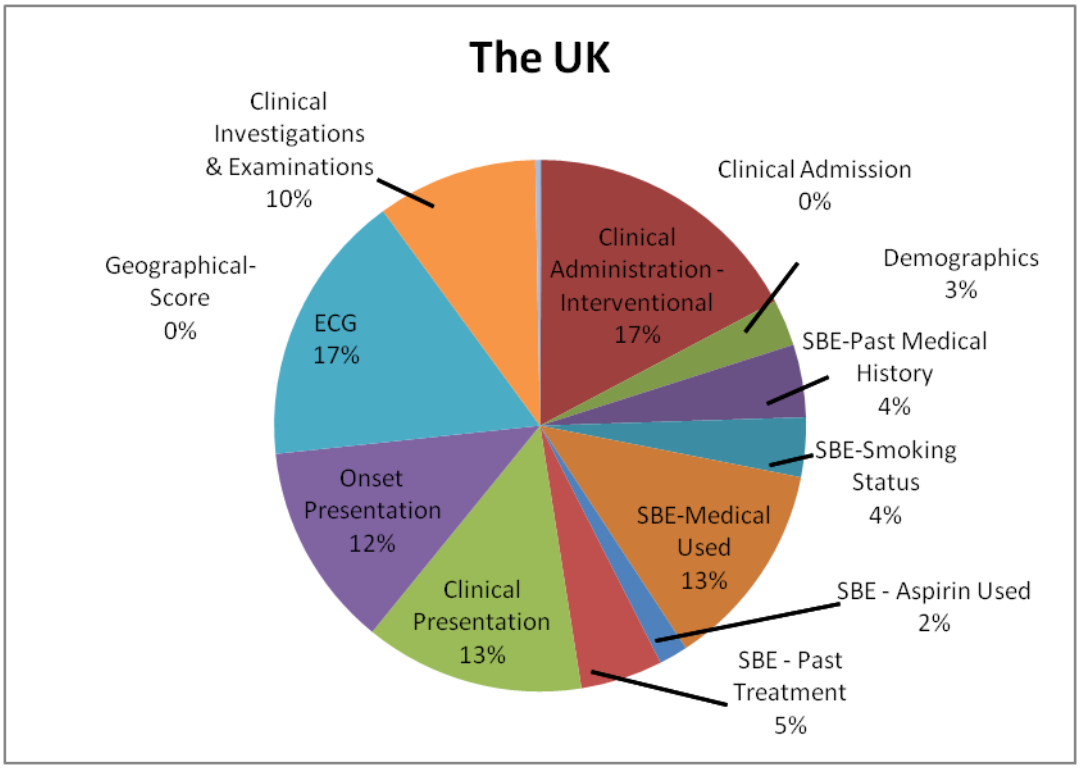


Figure 8: Mean percentage of missing values by clinical category-The UK dataset

Overall comparing Figure 7 and Figure 8, the UK datasets had more missing values when compared to the Malaysian dataset. These missing values were dominated by attributes in clinical investigations and examinations, clinical presentation, and medical status before the event for medical used categories. Meanwhile, for the Malaysian dataset, a high percentage of missing values were detected for status before the event in terms of past medical history and aspirin use. On the other hand, the UK dataset also had notable missing values in attributes under the ECG category.

4.5. Methodology Review

This section describes the findings upon reviewing the model development using WEKA classification algorithms, effect of missing values, and evaluation of random sampling methods.

4.5.1 Prediction Modelling Using WEKA

A preliminary study was carried out, and the related paper was presented at the 33rd SGAI International Conference on Artificial Intelligence (2013) in Cambridge, England. The study evaluated the process of developing prediction models using WEKA. Hence, some 960 Malaysian patients were employed for model development of NB, DT, and MLP.

4.5.2 WEKA Classification Algorithms

The objective of this review was to determine which algorithms were 'unsuitable' for the datasets as part of the Objective 1 of the study. Therefore, in the next phase, the datasets were not trained using the 'unsuitable' algorithms. In addition, this serves as a basic guideline for other researchers working with datasets that have similar characteristics.

Three input datasets were generated for each Malaysian and UK dataset, as described in Table 5.

Table 5: Input datasets for review of classification algorithms

Input Datasets	Descriptions	No. of predictors (Malaysia/UK)
Baseline modelling datasets	These are datasets prepared for model development	75/65
Common datasets	Subsets of baseline modelling datasets that included only the common attributes found in both the Malaysian and UK datasets. Refer to Appendix A:A.4*The Common Datasets.*	18/18
AMIS datasets	Subsets of baseline modelling datasets that only employed attributes from the AMIS model(Kurz et al., 2009). Refer to Appendix A: A.5*Characteristics of AMIS Model Vs The UK and Malaysian datasets.*	6/5

The three input datasets as presented in Table 5 used for this particular task was to have a variation of input datasets as to look into how in general, each of the ML algorithm reflect the performance of the model even when the predictors is reduced. Baseline modelling datasets considered all the attributes - allow for broad range of possible predictors to predict the outcome, and common datasets and AMIS datasets are the datasets with pre-selected attributes. The attributes in common datasets were selected solely based on the common attribute in the Malaysia and the UK dataset, regardless of its importance to the outcome and without any clinical

reasoning. AMIS dataset on the other hand, presented the set of attributes that have been used to develop ACS model i.e. AMIS model. Thus, the set of predictors used has been significantly evaluated.

In addition, as mentioned previously, two thirds of each dataset was used for training, while the remaining one third was reserved for validation. The datasets were trained on 29 algorithms and validated. The performance of each model was examined using AUC. As such, an AUC score of 0.65 and below is considered unsuitable.

Four algorithms have obtained an AUC score of 0.5 for all input datasets; these algorithms, VP, CR, Ridor, and ZR, were thus considered ‘unsuitable.’ Further, the SVM, JRip, OneR, and BFT algorithms were also deemed 'unsuitable' as their AUC scores were consistently below 0.6. Finally, j48, j48Graft, SC, and KNN resulted in fluctuating AUC scores between the three input datasets. Nevertheless, each of these algorithms has an average AUC score of below 0.65 for the three input datasets. Hence, j48, j48graft, SC, and KNN were also considered to be 'unsuitable.' As a result, only 17 algorithms were found suitable for further evaluation and model development.

In addition, the study also noted that LWL, MLP, DTNB, BFT, and LMT required notably more time for training and validating a model. On the other hand, KNN, RF, RT and PART were notably prone to overfitting.

Detailed results can be found in Appendix B: B.1 *WEKA Classification Algorithms*.

4.5.3 Missing Values

Since the datasets consist of a large number of attributes and quite a large percentage of missing values, it was not possible to create large enough training sets with complete cases. Therefore, removing incomplete cases as a way of handling missing values was not possible. Hence, this study explored the possibility of removing attributes with missing values. As such, this exercise explored the effect of removing attributes with various percentages of missing values.

The baseline datasets were employed in this exercise. Another five sets of input datasets were formed from the baseline datasets. They were:-

1) Baseline datasets with no missing values (BD_No_Mssg) - Removed all attributes with missing values.
2) Baseline datasets with 5% missing values (BD_5Prct_Mssg) - Removed all attributes with more than 5% missing values.
3) Baseline datasets with 10% missing values (BD_10Prct_Mssg) - Removed all attributes with more than 10% missing values.
4) Baseline datasets with 15% missing values (BD_15Prct_Mssg) - Removed all attributes with more than 15% missing values.
5) Baseline datasets with 20% missing values (BD_20Prct_Mssg) - Removed all attributes with more than 20% missing values.

All the datasets were split into training (2/3) and validation (1/3) sets. The training sets were used to develop the models using 17 algorithms, and the validation sets were used to validate the model. The AUCs of each model were then compared and analysed.

Collectively, the results suggest that removing all attributes with missing values resulted in poor model performance. The performance of the models started to improve when more attributes with a larger percentage of missing values were included in the datasets. Generally, most of the algorithms in both the Malaysian and UK datasets showed better AUC for the BD_15Prct_Mssg and BD_20Prct_Mssg datasets. This result suggests the possibility that most of the attributes with 15-20% missing values are indeed important predictors for the model. Thus, it does not appear to be wise to remove attributes with missing values, as this might remove important predictors. As a result, it was decided to first execute the feature selection tasks with all the attributes, including those with missing values. Once the attributes were reduced, the incomplete cases were handled accordingly.

Detailed results can be found in Appendix B: *B.2.Missing Values*

4.5.4 Random Sampling

The objective of this section was to determine and confirm the best proportion of training and validation sets for the random sampling method. Generally, two thirds for training is a common practice with which to randomly split datasets. In this section, nevertheless, different random splits were applied to the training and validation sets. The input datasets used for the exercise were the baseline modelling datasets. Thus, the input datasets were randomly divided into 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90 proportions for training and validation, respectively. These datasets were trained on 29 algorithms.

No pattern emerged that suggested that a specific proportion of random splitting was good for both datasets and all algorithms. The AUC of similar algorithms produced good results with varied percentages of random splitting when different datasets were applied. However, the results showed that a range of 70:30 to 40:60 splits yielded convincing AUC results on almost all classifiers run on both the Malaysian and UK datasets. As such, following the standard practice, this study opted to randomly divide the datasets into a2/3 split for model development.

4.6. Discussion and Conclusion

This chapter introduced the derivation cohorts for the study, along with a summary of the datasets and populations' characteristics. As a result, the final outcome of this chapter has been the baseline datasets and baseline modelling datasets. Baseline datasets were used to annotate the population characteristics, and the baseline modelling datasets were used for model development in the study.

The cohorts originated from two different regions: Asian and Western. The Asian dataset was comprised of Malaysian patients as a whole, while the Western dataset contained only a specific part of the UK: Leeds. Although, the Leeds, UK dataset represents only part of the UK, the sample was assumed to reflect the whole of the UK as there were no obvious difference in most of the patients' characteristics between our sample and the whole of the UK, as studied by Gale et al. (2008). The Asian and

Western datasets resulted in different characteristics within the samples. Moreover, the datasets were composed mainly of standard information pertaining to ACS, as the registries were developed based on standard international registries, as well as guidelines for ACS and CVD.

Nevertheless, the sole limitation of these datasets has been their quality. This issue was expected as the data were derived originally from EHR. Although the data had been validated and cleaned before being transferred into the registry, the quality of the datasets continued to be an issue to be addressed. The quality issue resulted in excluding approximately 90% of the UK records. However, the number of records left in the UK dataset was sufficient for the development of a reliable prediction model using ML, as well as for validating the model (Mukherjee et al., 2003, Beleites et al., 2013). Moreover, existing ACS models, such as MACE, were developed from samples smaller than the UK dataset (Hu et al., 2016). The quality issue in the datasets also required that a hefty amount of time be spent preparing the datasets for model development. In addition, issues related to the quality of the data dictionary were time constraining, especially when comprehending the datasets. It is also important to note the risk in having to exclude large number of observations from model development when data quality is at stake.

From the 29 ML algorithms identified, only 17 displayed the potential to be exercised in model development. In addition, LWL, MLP, DTNB, BFT, and LMT were found to be time-consuming in developing models, whilst KNN, RF, RT, and PART were inclined to overfitting.

Furthermore, as the datasets consists of a large number of attributes and contain a large number of missing values, removing all the incomplete cases from the datasets was not viewed as a strategic way of handling the missing values. However, removing attributes that hold the missing values was also not deemed a worthy strategy. The investigation into removing attributes that contained a certain percentage of missing values revealed that, as more attributes were removed (to minimised missing values of the datasets), the performance tended to degrade. This result indicates that most of attributes with missing values are indeed important predictors.

Hence, the study retained all attributes and missing values of the baseline modelling datasets when identifying a set of predictors for model development (feature selection). Only after the feature selection had been executed for model development, appropriate measures were applied to handle the missing values.

Moreover, the findings from reviewing varied proportions of hold-out random sampling were consistent with the standard hold-out strategy, i.e., the 2/3 distribution. In addition, the study also found that a range between a 70:30 split and 40:60 split yielded convincing AUC results. Therefore, this study had decided to use a hold-out random sampling method with a2/3 distribution to evaluate the developed models.

The following chapter demonstrates the implementation of feature selection methods to generate the input datasets for model development.

Chapter 5: Feature Selection and Model Development

This chapter describes the process of evaluating various subsets of predictors to fulfil Objective 2 of the study. As a result, varied sets of predictors were identified and were further used as input datasets for model development. The predictive performance of the developed models are also analysed and presented in this chapter.

5.1. Method

In the data preparation stage, attributes from raw datasets were thoroughly scanned and analysed to choose potential predictors that met the objectives of developing ACS prediction models. With that, all attributes that were duplicates, database-related, unknown, irrelevant, non-standardised, and dependent (functioned as additional attribute to cater missing values),have been excluded from the datasets. However, even after eliminating some of these attributes, the remaining number of attributes was still considered massive as there were more than 50 attributes. Therefore, to construct a simpler model with good predictive power, feature selection was applied in the model development process.

As such, in selecting sets of predictors for model development, various ML automated feature selection methods were evaluated. The potency of predictors from existing ACS models that were appropriate for adoption in developing simplified and customized prediction models was also assessed. In addition, the strength of predictors from different clinical categories in producing good models was also investigated. Sets of predictors from these three outcomes were then employed as input datasets for model development. The prediction models were then built on 17 classification algorithms. The best models for each of the Malaysian and UK datasets were chosen by comparing the AUC scores obtained by using the validation datasets.

5.1.1 Handling Missing Values

In the methodology review, specifically Section4.5.3 on missing values, it was decided to execute feature selection with all attributes, regardless of missing values, and, later, to remove the incomplete cases before model development. For future note, the amount of training and testing data was decreased when incomplete cases were omitted. Nonetheless, some incomplete cases could not be dismissed as they considerably reduced the number of training cases. As a result, for such cases, all the missing values were left for the algorithms to handle, as each selected algorithm has a method for handling missing values(Su et al., 2008).

5.1.2 Evaluating Automated ML Feature Selection

Two main methods from WEKA were employed to assess and identify subsets of attributes from both the Malaysian and UK datasets. As the study evaluated a range of ML algorithms suitable for the datasets, selecting a feature selection method that runs together with model development process could have been tricky and intricate. Thus, the chosen feature selection methods evaluated in the study were implemented as pre-processing procedures, i.e., before the model development process. The feature selection methods employed were the *subset* and *wrapper* methods. The *subset* method refers to a type of *filter* method that does not depend on any classification algorithm. This method selects subsets of attributes during the pre-processing steps before running the dataset into any classifier algorithm. Two types of *filter* methods are: 1) *univariate-filter*, and 2) *multivariate-filter (subset)* methods. The study adopted the latter method because it considers the relationship of individual attributes, as well as the correlation between attributes towards the outcome.

The subset methods applied in the study were: 1) *Correlation-Based-Feature-Selection (CFS)*, and 2) *FilterSubset. CFS* identifies a subset of attributes by selecting attributes with high correlation with the class, yet low correlation with each other (Hall and Smith, 1998). In WEKA, this subset method is known as *CfsSubsetEval*. On the other hand, the *FilterSubset*

employs a similar technique to *CFS*, except that, in selecting the attributes, *FilterSubset* divides a dataset into subsamples.

The *wrapper* method (Kohavi and John, 1997) incorporates a learning algorithm in the selection of attributes. As in WEKA, the subset evaluator will first detect all the possible subsets of attributes within the dataset. Next, these sets of attributes are trained on a specified classifier algorithm using a cross-validation technique. The best set of attributes is the one that performed the best on the clarification algorithm. As such, this study adopted two classification algorithms for the *wrapper* method, which were NB and LG. NB and LG retained their exceptional predictive performances in the prior task(refer to the Section4.5.2), in which the average AUC scores for these two classification algorithms exceeded 0.7.

In WEKA, the feature selection method implements a specific search method to determine a set of attributes. The search method utilized in this study was the *Greedy search* strategy using the *forward selection* approach. The *Greedy search* strategy forms a subset of attributes by progressively adding attributes to the subset until the best subset appears. It has been claimed that the *Greedy search* is computationally advantageous and robust against overfitting (Guyon and Elisseeff, 2003).

5.1.3 Evaluating Predictors of Existing ACS Models

Sets of predictors were manually select based on the attributes of the existing ACS prediction models. The ACS prediction models referred to in this task were: 1) TIMI(Antman et al., 2000),2) PURSUIT(Boersma et al., 2000), 3) Grace (In-hospital) (Granger et al., 2003), 4) GUSTO-I(Lee et al., 1995), 5) AMIS(Kurz et al., 2009), 6) Serbia(Sladojević et al., 2015), 7) C-ACS (Huynh et al., 2013), 8) EMMACE (Dorsch et al., 2001), and 9) MACE (Hu et al., 2016). The purpose of this task was to evaluate the potency of the existing sets of predictors to be adopted in building customized prediction models on other cohorts (e.g., the Malaysian and UK cohorts)using ML algorithms. A set of predictors was selected from the combination of predictors from the selected nine ACS prediction models. These predictors were then matched with the attributes available in the Malaysian and UK datasets.

Sets of predictors were also extracted from each of the seven selected ACS models, which were: 1) AMIS, 2) EMMACE,3) Canada ACS Risk Score, 4) GRACE, 5) PURSUIT, 6) GUSTO-I, and 7) Serbia.

5.1.4 Evaluating Predictors of Different Clinical Categories

In the data processing phase, the datasets were categorised based on clinical reasoning. By categorising these datasets, the clinical data were grouped into similar or related items/events for better visualisation and understanding of the datasets. The attributes were grouped into id, demographics, status before event (medical history and medication pre-admissions), clinical presentation, ECG, and baseline investigations. The purpose of selecting predictors based on clinical categories is to evaluate the impact predictors from each clinical category had in constructing good prediction models. Another reason was to assess the effect of having predictors concerning medication taken before admission as part of the predictors..

Thus, sets of predictors were grouped into five combinations of clinical categories, which were:-

CATA1 - demographics and medical history

CATA2 - demographics, medical history, and medication pre-admissions

CATA3 - demographics,medical history,medication pre-admissions, and clinical presentation

CATA4 - demographics, medical history,medication pre-admissions, clinical presentation, and ECG

CATA5 - demographics, medical history,medication pre-admissions, clinical presentation, ECG, and baseline investigations

Each combination has its own set of predictors, for which the models were then developed.

Another set of predictors was also formed using the same approach, but also applying the *CFS* feature selection method. Predictors in each clinical category were filtered using the *CFS* feature selection method, and then all the filtered predictors from each category were combined to form

CATA7. CATA7 was formed to examine the benefits of having fewer predictors when using ML feature selection in evaluating predictors from different clinical categories.

5.2. Results

This section presents the sets of predictors resulting from the three tasks described in Sections 5.1.2 - 5.1.4. It also demonstrates the results of models developed based on the established sets of predictors from these three tasks. The best prediction models for both the Malaysian and UK datasets were examined based on their discrimination capability using AUC. The best prediction models also subsequently represent the best subsets of predictors for each of the datasets. Lastly, the final part of this section describes the performances of the models built by ML algorithms.

5.2.1 Evaluating Automated ML Feature Selection: Sets of Predictors

The subsets of predictors chosen by applying *CFS*, *FilterSubset*, *wrapper* with LG algorithm (*WrapperLG)*, and *wrapper* with NB algorithm (*WrapperNB)* methods on the Malaysian and the UK datasets are tabulated in Table 6.

Table 6: Subsets of predictors selected by ML automated feature selection

Automated feature selection methods	Malaysian	The UK
	List of predictors	List of predictors
CFS	1) ptageatnotification 2) heartrate 3) bpsys 4)bpdias 5) ecgabnormtypetwave* 6) lvef	1) Age.At.Admission 2) X224.Beta.Blocker* 3) X220.Systolic.BP 4) X314.Where.cardiac.arrest* 5) X424.Reinfarction*
FilterSubset	1) ptageatnotification 2) heartrate 3) bpsys 4)bpdias	1) X314.Where.cardiac.arrest*
WrapperLG	1) sdpid* 2) ptageatnotification 3) cdm* 4) chpt * 5) canginamt2wk* 6) weight 7) ecgabnormlocationrv* ~~8) lvef~~	1) ADMISSION_YEAR* 2) Age.At.Admission 3) X204.Where.Aspirin.Given* 4) X314.Where.cardiac.arrest* ~~5) Clopidogrel*~~ ~~6) X228.Glucose*~~ ~~7) X315.Presenting.Rhythm*~~ ~~8) X236.Site.of.Infarction*~~
WrapperNB	1) chpt* ~~2) ccap*~~ 3) canginamt2wk* ~~4) cpvascular*~~ 5) CNONE* 6) ACS_SYMPTOMS_BEFORE_ ADMISSION* ~~7) weight~~ ~~8) waistcircumf~~ 9) ecgabnormtypenonspecific* 10) ecgabnormlocationll* 11) ecgabnormlocationtp* 12) ecgabnormlocationrv* ~~13) tc~~ ~~14) hdlc~~ ~~15) ldlc~~ ~~16) tg~~ 17) asapre* 18) adpapre* 19) gpripre* 20) heparinpre* 21) lmwhpre* 22) bbpre* 23) aceipre* 24) arbpre* 25) statinpre* 26) lipidlapre* 27) diureticpre* 28)calcantagonistpre* 29) oralhypoglypre* 30) insulinpre* 31) antiarrpre*	1) ADMISSION_YEAR* 2) ADMISORC* 3) X222.Admitting.Consultant* 4) Age.At.Admission 5) X209.Peripheral.Vascular.Disease* 6) X211.Asthma.or.COPD* 7) X218.Previous.PCI* 8) X219.Previous.CABG* 9)ONSET_SYMPTOMS_BEFORE_ADMISSION* 10) X314.Where.cardiac.arrest* 11) ATTEND_NON_INTERVENTIONAL_ HOSPITAL* ~~12) X238.Thienopyridine.inhibitor.use*~~ ~~13) X220.Systolic.BP~~ ~~14) X230.Weight~~ ~~15) X315.Presenting.Rhythm*~~ ~~16) X424.Reinfarction*~~ ~~17) X237.ECG.QRS.Complex.duration*~~ ~~18) X236.Site.of.Infarction*~~ ~~19) X215.Cholesterol~~ ~~20) X231.LVEF~~ ~~21) X347.Assess.at.non.intevention.hospital*~~

The underlined predictors represent the predictors from the category of medication received before admission; The strikethrough predictors illustrate the discarded predictors due to missing values. The * denotes categorical predictors.

Considering Table 6, the *WrapperNB* method has the most predictors. The method identified 31 and 21 attributes as potential predictors for the Malaysian and UK datasets, respectively. In contrast, *FilterSubset*reduced the attributes to only one potential predictor for the UK dataset.

The potential predictors identified in all ML feature selection methods reflecta mix of varied clinical categories of clinical admission and demographics, past medical history, medication received before admission, clinical presentation, ECG, and clinical investigation. Age was selected by all the methods, with the exception of the*WrapperNB* method; this suggests that age is the most essential predictor for the datasets. In addition, predictors under clinical presentation (e.g., SBP, heart rate, and presentation of cardiac arrest) and ECG categories emerged as the most selected predictors by the four ML feature selection methods. On the other hand, attributes from the category describingmedication received before admission were the least selected by most of the feature selection methods, with the exception of the*WrapperNB*method on the Malaysian dataset, for which almost all attributes in this category were selected. Hence, the effect of having predictors under the category of medication received before admission has been further investigated.

As described in Section 5.1.1, to handle missing values, only complete cases were considered for model development. Therefore, the missing cases of the datasets from applying *CFS* and *FilterSubset*were excluded before model development. Nonetheless, the approach to addressing the missing values was different for the datasets when applying the *wrapper* methods. This is because the *wrapper* method extracted a considerably substantial number of predictors with large percentages of missing values. Thus, excluding the incomplete cases from the datasets applying the *wrapper* methods was not practical as it reduced a large number of cases for model development. Therefore, the predictors selected by the *wrapper* method were further filtered by discarding predictors with missing values >20%. As a result, for the Malaysian dataset, one predictor of the eight predictors selected by the *WrapperLG* method and eight predictors from the 31 predictors selected by the *WrapperNB* method were dismissed. Filtering the missing values on the UK dataset, four attributes were excluded

from set of predictors selected by the *WrapperLG* method, while only 11 predictors remained on the list of predictors selected by the *WrapperNB* method. All the predictors that had been removed are represented as strikethrough in Table 6. After applying the strategy in handling missing values for the datasets, the reduced sample size for model development is illustrated in Figure 9.

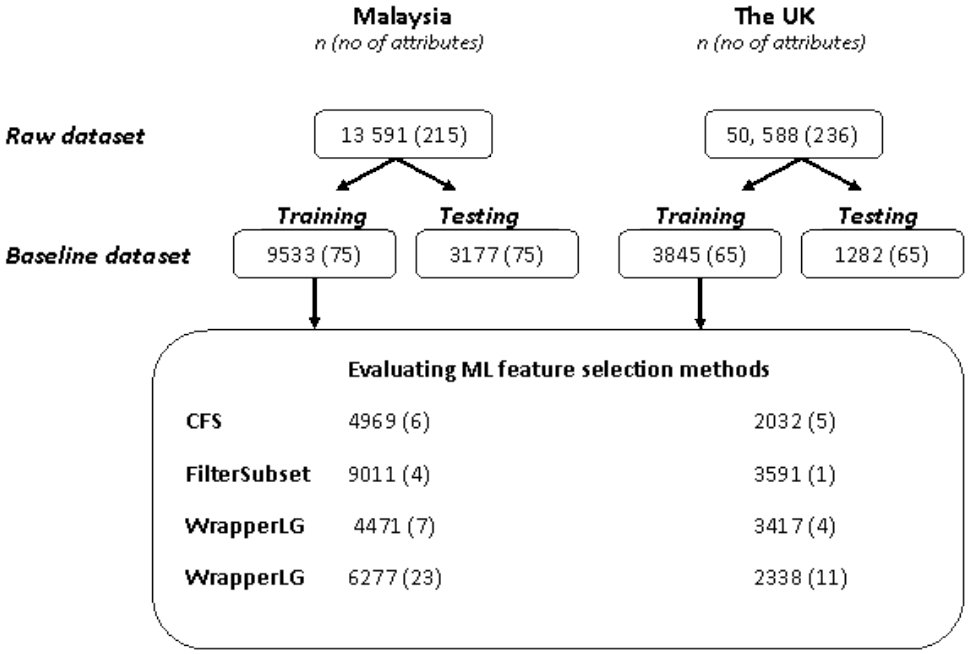


Figure 9 : Sample size for model development of evaluating ML feature selection method

5.2.2 Evaluating Automated ML Feature Selection: The Prediction Models

Tables 7and 8 tabulate the performances of the models developed based on sets of predictors extracted from ML automated feature selection methods for the Malaysian and UK datasets, respectively.

Table 7: The Malaysian models developed based on sets of predictors extracted from ML automated feature selection methods

Models	CFS_MY	FS_MY	WR_LG_MY	WR_NB_MY
BN	0.762	0.765	0.640	0.615
NB	**0.794**	0.765	0.667	0.609
LG	**0.801**	**0.777**	0.676	0.605
MLP	0.773	**0.776**	0.558	0.549
LWL	0.500	0.723	0.660	0.524
DT	0.500	0.768	0.500	0.500
DTNB	0.500	0.768	0.500	0.500
PART	0.702	0.616	0.673	0.513
ADT	0.773	0.768	**0.769**	0.569
DS	0.645	0.613	0.626	0.532
FT	0.500	0.500	0.500	0.500
LT	0.774	**0.787**	**0.708**	0.576
LMT	**0.802**	**0.776**	0.500	0.500
NBT	0.762	0.767	0.640	0.619
RF	0.758	0.753	**0.731**	0.570
RT	0.533	0.544	0.533	0.499
REPT	0.500	0.616	0.679	0.500

The underlined values represent AUC>0.8; the blue-coloured values indicates the three best models; the grey-shaded attribute denotes the overall best model

In Table 7, CFS_MY, FS_MY, WR_LG_MY, and WR_NB_MY represent models developed for the Malaysian dataset using *CFS*, *FilterSubset,* and *wrapper* methods (*WrapperLG* and *WrapperNB*), respectively.

The results in Table 12 generally suggest that the models developed based on predictors from *subset* methods (*CFS* and *FilterSubset*) generated better AUC scores in comparison to the *wrapper* methods (*WrapperLG* and WrapperNB). In fact, the average AUC scores for WR_LG_MY and WR_LG_MY models were 0.621 and 0.546, respectively. Moreover, the lowest AUC score produced by the WR_NB_MY model may have resulted from loss of several essential predictors due to removal of attributes in addressing missing values. The results also showed that having predictors from the category of medication received before admission(selected by WrapperNB) did not improve the models.

Comparing the models developed by predictors of *subset* methods, generally, CFS_MY performed better than FS_MY. In developing CFS_MY

models, two classification algorithms attained an AUC > 0.8, i.e., LG (AUC = 0.801) and LMT (0.802). Additionally, NB also displayed a considerably good AUC score (0.794) in building the CFS_MY model. On the other hand, the results demonstrated that LT (AUC = 0.787), LG (AUC = 0.777), LMT (AUC = 0.776), and MLP (AUC = 0.776) emerged as the best algorithms for the FS_MY models.

Table 8: The UK models developed based on sets of predictors extracted from ML automated feature selection methods

Models	CFS_UK	FS_UK	WR_LG_UK	WR_NB_UK
BN	0.793	0.668	0.789	0.824
NB	**0.889**	0.595	**0.818**	**0.847**
LG	**0.869**	**0.706**	**0.809**	**0.847**
MLP	**0.859**	0.609	0.765	0.773
LWL	0.815	**0.702**	0.741	0.795
DT	0.709	0.668	0.798	0.700
DTNB	0.714	0.668	0.781	0.817
PART	0.653	0.621	0.779	0.758
ADT	0.794	0.656	0.807	0.805
DS	0.628	0.664	0.669	0.695
FT	0.850	**0.706**	0.590	0.635
LT	0.764	0.689	0.780	0.844
LMT	0.868	0.667	**0.807**	**0.867**
NBT	0.791	0.668	0.795	0.822
RF	0.713	0.646	0.800	0.778
RT	0.603	0.583	0.657	0.618
REPT	0.500	0.651	0.676	0.699

The underlined values represent AUC>0.8; the blue-coloured values indicate the three best models; the grey-shaded attribute denotes the overall best model

In Table 8, CFS_UK, FS_ UK, WR_LG_ UK, and WR_NB_ UK represent models developed for the UK dataset using *CFS*, *FilterSubset,* and *wrapper* methods (*WrapperLG* and *WrapperNB*), respectively.

As in the Malaysian dataset, CFS_UK also exhibited the best performance for the UK dataset (Table 8) in comparison to other automated feature selection methods by obtaining the highest AUC for NB (AUC =0.889). To the contrary, models developed with predictors chosen by *FilterSubset* had the lowest AUC score in most of the algorithms. In fact, the models developed with predictors selected by the *wrapper* methods

achieved noticeably good AUC scores, with more than half of the algorithms hitting AUC> 0.75. Among the best algorithms that achieved AUC > 0.8 in CFS_UK, WR_LG_UK, and WR_NB_UK were NB, LG, and LMT.

In conclusion, for both the Malaysian and UK datasets, *CFS* methods appeared to be the best feature selection methods. *FilterSubset* also appeared to be a good feature selection method for the Malaysian dataset, unlike the *wrapper* methods. Unfortunately, the results showed otherwise for the UK dataset, in which *wrapper* method seemed to produce better models than the *FilterSubset* method. The results from both tables also indicate biasedness on models developed using the same algorithm used by *wrapper* feature selection methods (i.e., NB and LG).Nevertheless, the cases of biasedness were not too vivid.

5.2.3 Evaluating Predictors of Existing ACS Models : Sets of Predictors

Table 9 presents the sets of predictors derived from extracting predictors from ACS models.

Table 9: Set of predictors of existing ACS models

Referred ACS Models	Malaysian List of predictors	The UK List of predictors
All_LR	1) ptsex* 2) ptageatnotification 3) statusaspirinuse* 4) cheartfail* 5) ACS_SYMPTOMS_BEFORE_ADMISSION* 6) heartrate 7) bpsys 8) bpdias 9) killipclass* 10) lvef	1) Age.At.Admission 2) X107_Gender* 3) X213.Heart.Failure* 4) X204.Where.Aspirin.Given* 5) X220.Systolic.BP 6) X221.Heart.Rate 7) X314.Where.cardiac.arrest* 8) ST_Segment_Deviation* ***** Attributes lvef was deleted due To large amount (94%) of missing values***
AMIS	1) ptageatnotification 2) cheartfail* 3) ccerebrovascular* 4) heartrate 5) bpsys 6) killipclass* ***** No Pre-hospital cardiopulmonary resuscitation attribute***	1) Age.At.Admission 2) X210.Cerebrovascular.Disease* 3) X213.Heart.Failure* 4) X220.Systolic.BP 5) X221.Heart.Rate ***** No Pre-hospital cardiopulmonary resuscitation attribute***
EMMACE	1) ptageatnotification 2) heartrate 3) bpsys	1) Age.At.Admission 2) X220.Systolic.BP 3) X221.Heart.Rate
C-ACS	1) ptageatnotification 2) heartrate 3) bpsys 4) killipclass*	***** Same as EMMACE ** No killip class attribute***
GRACE	1) ptageatnotification 2) heartrate 3) bpsys 4) killipclass* 5) ACS_SYMPTOMS_BEFORE_ADMISSION* ***** No serum creatinine attribute*** ***** No positive initial cardiac enzyme attribute***	1) Age.At.Admission 2) X221.Heart.Rate 3) X220.Systolic.BP 4) X314.Where.cardiac.arrest* ***** No killip class attribute*** ***** No serum creatinine attribute*** ***** No positive initial cardiac enzyme attribute***
PURSUIT	1) ptageatnotification 2) gender* 3) heartrate 4) bpsys 5) st_segment depression* ***** No sign of heart failure attribute*** ***** No cardiac enzyme attribute***	1) Age.At.Admission 2) X107_Gender* 3) X221.Heart.Rate 4) X220.Systolic.BP 5) st_segment depression* ***** No sign of heart failure attribute*** ***** No cardiac enzyme attribute***
GUSTO-i	***** Same as C-ACS*** ***** No anterior infraction attribute***	***** Same as EMMACE*** ***** No killip class attribute*** ***** No killip class attribute*** ***** No anterior infraction attribute***
Serbia	1) Ptageatnotification 2) Bpsys 3) Heartrate 4) Bpdias 5) Lvef ***** No troponin value attribute***	1) Age.At.Admission 2) X220.Systolic.BP 3) X221.Heart.Rate 4) X231.LVEF ***** No DBP attribute*** ***** No troponin value attribute***

The * denotes categorical predictors

Referring to Table 9, the sets of predictors that matched the combination of predictors of the nine existing ACS prediction models are represented by All_LR. Ten predictors from the Malaysian dataset, and eight predictors from the UK dataset matched the combined set of predictors from the nine ACS prediction models (All_LR). The extra predictor for the Malaysian dataset was the *Killip class*. For the UK dataset, nine predictors were initially matched, but the *lvef* predictor was discarded due an enormous number of missing values(94%). Predictors with a large number of missing values could lead to misleading conclusions from a prediction model. Hence, the final set of predictors matching the UK dataset consisted of just eight predictors.

Meanwhile, the other sets of predictors that were drawn from each of the seven ACS models were represented by the names of the ACS models, i.e. AMIS, EMMACE, C-ACS, GRACE, PURSUIT, GUSTO-I, and Serbia.

As described in Section 5.1.1, in order to handle missing values, only complete cases were considered for model development. Thus, all the incomplete cases were removed, and the final sample size for model development of each input datasets is illustrated in Figure 10.

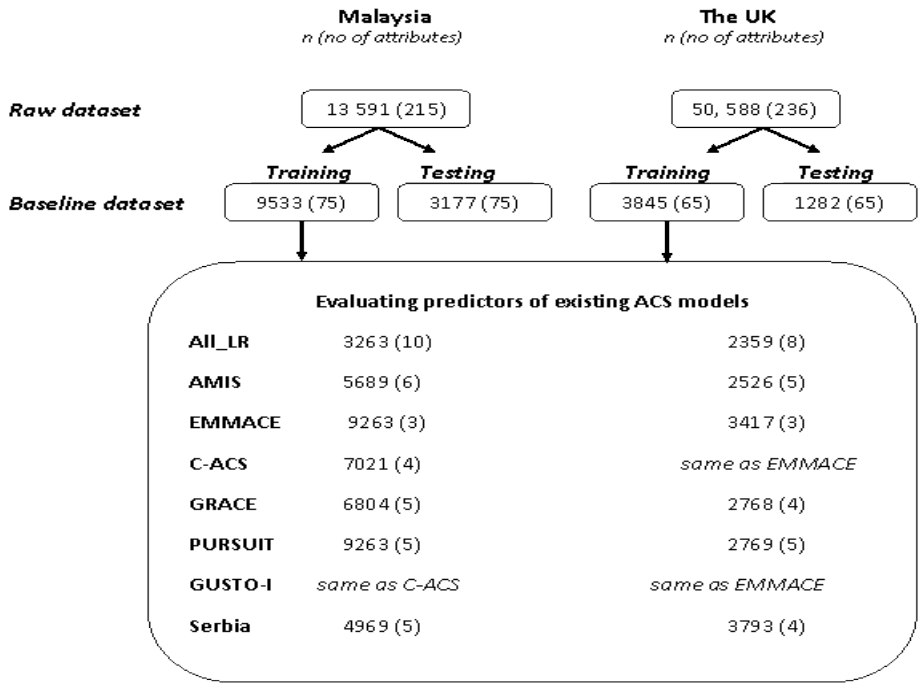


Figure 10: Sample size for model development for evaluating predictors of existing ACS models

5.2.4 Evaluating Predictors of Existing ACS Models: The Prediction Models

The performance of models constructed with sets of predictors adapted from existing ACS models are presented in Tables 9 and 10. Table 9 represents the results of the Malaysian models, while Table 10 presents the results of the UK models.

Table 9: The performance of Malaysian models adopting predictors from existing ACS models

Models	All_LR_MY	AMIS_MY	EMMACE_MY	C-ACS_MY	GRACE_MY	PURSUIT _MY	Serbia_MY
BN	**0.843**	0.815	0.749	0.813	0.844	0.789	0.763
NB	**0.845**	**0.829**	**0.775**	**0.822**	**0.906**	**0.836**	0.795
LG	0.842	0.827	0.773	0.822	**0.904**	**0.826**	**0.802**
MLP	0.767	0.792	0.771	0.814	0.890	0.811	0.793
LWL	0.792	0.806	0.729	0.780	0.858	0.741	0.705
DT	0.500	0.785	0.768	0.795	0.774	0.500	0.500
DTNB	0.773	0.821	0.768	0.823	0.774	0.500	0.790
PART	0.669	0.719	0.732	0.730	0.596	0.762	0.674
ADT	0.807	0.811	0.757	0.794	0.789	0.787	0.773
DS	0.647	0.593	0.611	0.603	0.684	0.696	0.645
FT	0.606	0.827	0.500	0.822	**0.902**	0.500	**0.802**
LT	0.811	**0.832**	**0.783**	**0.824**	0.823	**0.804**	0.774
LMT	**0.861**	**0.829**	**0.782**	**0.823**	0.868	0.500	**0.802**
NBT	0.768	0.815	0.749	0.813	0.846	0.789	0.763
RF	0.821	0.806	0.737	0.778	0.828	0.790	0.775
RT	0.563	0.568	0.542	0.574	0.644	0.549	0.607
REPT	0.761	0.714	0.717	0.721	0.829	0.500	0.500

The underlined values represent AUC>0.8; the blue-coloured values indicate the three best models; the grey-shaded attribute denotes the overall best model

In Table 9, All_LR_MY, AMIS_MY, EMMACE_MY, C-ACS_MY, GRACE_MY, PURSUIT_MY, and Serbia_MY represent models developed for the Malaysian dataset using predictors from a combination of nine ACS models, and predictors from each of AMIS, EMMACE, C-ACS, GRACE, PURSUIT, and Serbia models, respectively.

The overall results of Table 10for the Malaysian dataset indicate that the best AUC rate was achieved by adapting predictors from the GRACE model, in which three classification algorithms achieved AUC > 0.9, i.e., NB (AUC=0.906), LG (AUC=0.904), and FT (AUC=0.902). AMIS_MY, C-

ACS_MY, PURSUIT_MY, and Serbia_MY also produced fairly good AUC scores, for which at least the best three algorithms of each model achieved an AUC score >0.8. Although EMMACE_MY seemed not to perform as well as the other models (AUC>0.8), it still obtained an AUC rate of >0.75 in several algorithms, such as NB (AUC=0.775), LT (AUC=0.783), and LMT (AUC=0.782).

Table 10: The performance of the UK models adopting predictors from existing ACS models

Models	All_LR_UK	AMIS_UK	EMMACE_UK	GRACE_UK	PURSUIT_UK	Serbia_UK
BN	0.872	0.709	0.769	0.818	0.749	0.723
NB	**0.917**	**0.819**	**0.829**	0.826	0.770	**0.783**
LG	**0.874**	**0.844**	**0.833**	**0.827**	**0.775**	**0.782**
MLP	0.855	0.816	**0.819**	0.810	0.771	**0.787**
LWL	0.850	0.812	0.789	0.783	0.719	0.739
DT	0.693	0.500	0.500	0.799	0.731	0.500
DTNB	0.850	0.500	0.500	**0.830**	0.731	0.500
PART	0.772	0.643	0.500	0.717	0.752	0.500
ADT	0.799	0.713	0.793	0.786	0.752	0.736
DS	0.690	0.718	0.700	0.606	0.611	0.689
FT	**0.902**	0.500	0.500	**0.827**	0.549	0.500
LT	0.836	**0.818**	0.804	**0.831**	**0.783**	0.735
LMT	0.639	0.500	0.500	0.826	**0.772**	0.500
NBT	0.870	0.709	0.769	0.818	0.749	0.752
RF	0.840	0.779	0.769	0.782	0.751	0.704
RT	0.656	0.558	0.535	0.571	0.571	0.621
REPT	0.696	0.500	0.500	0.729	0.717	0.500

The underlined values represent AUC>0.8; the blue-coloured values indicate the three best models; the grey-shaded attribute denotes the overall best model

In Table 10, All_LR_UK, AMIS_UK, EMMACE_UK, GRACE_UK, PURSUIT_UK, and Serbia_UK represent models developed for the UK dataset using predictors from a combination of nine ACS models, and predictors from each of AMIS, EMMACE, GRACE, PURSUIT, and Serbia models, respectively.

Table 10 presents results of the UK models. All_LR_UK displayed outstanding results with two classifiers obtaining AUC>0.9. In addition, another seven algorithms scored AUC>0.8 on All_LR_UK. On the other hand, none of the algorithms obtained AUC>0.8 for PURSUIT_UK and

Serbia_UK. Nevertheless, some of these algorithms on PURSUIT_UK and Serbia_UK did gain AUC >0.75. All in all, a numbers of algorithms achieved AUC >0.8 on AMIS_UK, EMMACE_UK, and GRACE_UK.

Furthermore, the performance of the models was then compared with the c-statistics of the existing ACS models. Table 11 presents the comparison between the best models derived from this task and published c-statistics/AUC of existing ACS models.

Table 11: Comparison of predictive performance of the developed models and the existing ACS models

Prediction Models	Published C-Statistics/AUC	Malaysian	The UK
AMIS	0.875 (AODE)	0.832 (LT)	0.844 (LG)
EMMACE	0.76(Multivariable logistic regression)	0.783 (LT)	0.833 (LG)
C-ACS	0.75 (Multivariable logistic regression)	0.824 (LT)	0.833 (LG)
GUSTO-I	N/A (Logistic multiple regression)	0.824 (LT)	0.833 (LG)
GRACE In-Hospital	0.83 (Multivariable logistic regression)	0.906 (NB)	0.831 (LT)
PURSUIT	0.81 (death only) (Multivariable logistic regression)	0.836 (NB)	0.783 (LT)
Serbia	0.91 (ADT)	0.802 (LG)	0.787 (MLP)

The underlined values represent the AUC of developed model and gained better performance than the existing model

Comparing the AUC results obtained from this exercise with those of the published c-statistics/AUC of the seven existing ACS models as demonstrates in Table 11, the models constructed in this study displayed better performance than most of the five existing ACS models developed with traditional LG. In fact, almost half of the ML algorithms used in the study performed exceptionally better than the current models developed using traditional LG. As such, the results suggest that ML algorithms possessed the ability to construct better ACS prediction models. Nevertheless, models developed by adapting predictors from the AMIS and Serbia models(AMIS_MY, AMIS_UK, Serbia_MY, and Serbia_UK)showed lower predictive power than the AUC of the original AMIS and Serbia models. The AMIS and Serbia models were developed using ML algorithms, namely AODE and ADT, respectively.

The best models for the Malaysian dataset were developed by adapting predictors from the GRACE model, while the UK dataset achieved better models by adopting predictors of the combination of nine existing ACS models (All_LR_UK). To the contrary, employing EMMACE predictors resulted in the worst models for the UK dataset, as no algorithm achieved AUC>0.8. A similar scenario was noted when adapting predictors of the PURSUIT and Serbia models for the UK datasets.

5.2.5 Evaluating Predictors of Different Clinical Categories: Sets of Predictors

Five subsets of predictors were extracted from five groups of combinations of clinical categories. The groups were CATA1, CATA2, CATA3, CATA4, and CATA7. The detailed predictors for each group are tabulated in Appendix C.1 *Set of Predictors by Combination of Clinical Categor*y. As illustrated in the table in Appendix C.1, the number of predictors grew as more predictors of different clinical categories were added. For the Malaysian dataset, CATA1 had 19 predictors, CATA2 had 34 predictors, CATA3 had 43 predictors, CATA4 had 54 predictors, and CATA5 had 60 predictors. In terms of the UK dataset, CATA1 had 18 predictors, CATA2 had 23 predictors, CATA3 had 32 predictors, CATA4 had 37 predictors, and CATA5 had 40 predictors. Excluding missing cases was not feasible as huge number of cases would need to have been removed. Thus, in this particular case, the missing values were handled by the learned algorithm and no exclusion of instances was removed. All of the instances allocated to the training set as reserved in the earlier chapter were used for constructing and evaluating the models for this particular case. A total of 9533 instances in the Malaysian dataset, and a total of 3793 instances in the UK dataset were used to develop models with predictors of different clinical categories.

As described in Section 5.1.4, another set of predictors used in evaluating predictors of different clinical categories is represented by CATA7. CATA7 was formed by filtering the predictors from each clinical category using the *CFS* method, and combined all the filtered predictors of each clinical category. Table 12 presents the predictors of CATA7 for the Malaysian and UK datasets.

Table 12: Subsets of predictors for CATA7

Malaysian		The UK	
List of predictors	**Clinical Category**	**List of predictors**	**Clinical Category**
1) yradmit	Demographic	1) Age.At.Admission	Demographic
2) ptageatnotification	Demographic	2) X210.Cerebrovascular.Disease	Medical history
3) cpremcvd	Medical history	3) X212.Chronic.Renal.Failure	Medical history
4) cheartfail	Medical history	4) X213.Heart.Failure	Medical history
5) clung	Medical history	5) X217.Diabetes	Medical history
6) crenal	Medical history	6) X216.Smoking.Status	Medical history
7) heartrate	Clinical presentation	7) X204.Where.Aspirin.Given	Medical history
8) bpsys	Clinical presentation	8) X224.Beta.Blocker	Medical received
9) bpdias	Clinical presentation	9) X220.Systolic.BP	Clinical presentation
10) ecgabnormtypetwave	ECG	10) X314.Where.cardiac.arrest	Clinical presentation
11) ecgabnormtypebbb	ECG	11) X424.Reinfarction	ECG
12) ecgabnormtypenonspecific	ECG	12) X203.ECG.Determining.Treatment	ECG
13) ecgabnormlocational	ECG	13) X337.Troponin.Assay	Baseline investigations
14) ecgabnormlocationrv	ECG		
15) ldlc	Baseline investigations		
16) fbg	Baseline investigations		
17) lvef	Baseline investigations		
18) lmwhpre	Medical received		
19) aceipre	Medical received		
20) diureticpre	Medical received		
21) antiarrpre	Medical received		

The * denotes categorical predictors

For the CATA7 group, only complete cases were considered for model development. Therefore, all instances with missing values were removed, and the final sample size for model development is depicted in Figure 11.

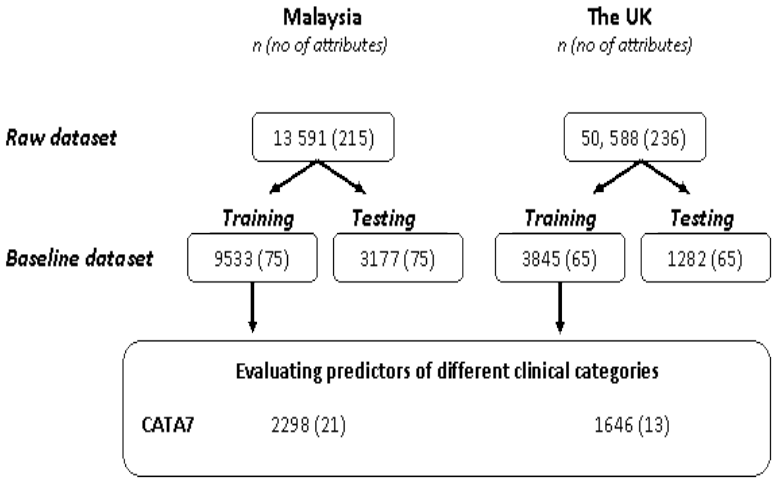


Figure 11: Sample size for model development for evaluating predictors from the CATA7 group

5.2.6 Evaluating Predictors of Different Clinical Categories: The Prediction Models

Tables13and 14indicatethe performances of the models with predictors of different clinical categories for the Malaysian and UK datasets, respectively.

Table 13: The Malaysian models with predictors of different clinical categories

Models	CATA1_MY	CATA2_MY	CATA3_MY	CATA4_MY	CATA5_MY	CATA7_MY
BN	**0.665**	**0.683**	0.760	**0.780**	0.779	**0.850**
NB	0.658	**0.679**	0.758	0.776	0.788	0.789
LG	**0.696**	**0.701**	**0.789**	**0.804**	**0.808**	**0.837**
MLP	0.599	0.606	0.758	0.762	0.798	0.647
LWL	**0.663**	0.665	0.748	0.753	0.756	0.699
DT	0.500	0.500	**0.783**	0.772	0.772	0.500
DTNB	0.465	0.499	0.479	0.490	0.490	0.750
PART	0.648	0.635	0.710	0.693	0.731	0.735
ADT	0.647	0.658	**0.795**	**0.795**	**0.807**	0.801
DS	0.618	0.618	0.680	0.680	0.680	0.677
FT	0.500	0.500	0.642	0.690	0.723	0.834
LT	0.668	0.662	0.735	0.740	0.740	0.743
LMT	0.500	0.500	0.767	0.768	0.790	**0.837**
NBT	0.655	0.675	0.654	0.628	0.798	0.569
RF	0.618	0.650	0.750	0.774	**0.799**	0.803
RT	0.598	0.594	0.678	0.640	0.720	0.515
REPT	0.618	0.622	0.663	0.659	0.734	0.500

The underlined values represent AUC>0.8; the blue-coloured values indicate the three best models; the grey-shaded attribute denotes the overall best model

In Table 13, CATA1_MY, CATA2_MY, CATA3_MY, CATA4_MY, CATA5_MY, and CATA5_MY represent models developed for the Malaysian dataset using the CATA1, CATA2, CATA3, CATA4, CATA5, and CATA7 groups, respectively.

Observing the results of Table 13,the model performances of CATA1_MY and CATA2_MY were unsatisfactory. This indicate that the combination predictors from the demographic, status before events, and medication received before admission categories failed to produce good ACS models. However, the performance of the models started to improve when predictors from the clinical presentation, ECG, and baseline investigation categories were included (i.e., CATA3_MY, CATA4_MY, and CATA5_MY). Nevertheless, only two algorithms managed to achieve AUC>0.8 in either CATA3_MY, CATA4_MY, or CATA5_MY, which were LG and ADT. In fact, the best three classification algorithms for CATA3_MY were LG, DT, and ADT with AUC scores of 0.789, 0.783, and 0.795,

respectively. Following the results obtained from CATA3_MY, ADT and LG remained as two out of the three best algorithms for CATA4_MY and CATA5_MY. Additionally, BN and RF emerged among the best three algorithms for CATA4_MY and CATA5_MY.

Meanwhile, overall better models (in most of the algorithms) were developed using predictors from CATA7 compared to CATA1_MY, CATA2_MY, CATA3_MY, CATA4_MY, and CATA5_MY. This might be due to applying the feature selection method, making the CATA7_MY simpler than the other models. However, surprisingly, for several algorithms, such as DT, NBT, RT, and REPT, CATA7_MY displayed a sharp plunge in their performances to AUC values around 0.500; indicating nil discriminatory ability. Nevertheless, other classification algorithms exhibited improvement, with six algorithms achieving AUC scores above 0.8. All in all, the best algorithms were BN (AUC = 0.850), LG (AUC = 0.837), and LMT (AUC = 0.837) for models with CATA7 predictors.

Table 14: The UK models with predictors of different clinical categories

Models	CATA1_UK	CATA2_UK	CATA3_UK	CATA4_UK	CATA5_UK	CATA7_UK
BN	**0.738**	0.687	0.812	**0.832**	**0.832**	0.808
NB	0.725	**0.738**	**0.832**	**0.848**	**0.847**	**0.866**
LG	**0.746**	**0.743**	0.817	0.820	0.818	0.742
MLP	0.637	0.762	0.804*	0.769	0.798	0.669
LWL	0.709	0.724	0.693	0.696	0.670	0.741
DT	0.500	0.500	0.668	0.664	0.664	0.598
DTNB	0.448	0.500	0.672	0.786	0.774	0.595
PART	0.700	0.500	0.812	0.798	0.804	0.810
ADT	0.726	0.735	0.821	**0.833**	**0.841**	**0.835**
DS	0.689	0.689	0.669	0.669	0.669	0.596
FT	0.592	0.500	**0.824**	0.822	0.822	0.565
LT	**0.741**	**0.752**	0.781	0.781	0.781	0.770
LMT	0.500	0.500	**0.824**	0.830	0.830	**0.888**
NBT	0.725	0.720	0.788	0.825	0.826	0.810
RF	0.677	0.623	0.802	0.818	0.812	0.693
RT	0.640	0.570	0.717	0.664	0.740	0.604
REPT	0.725	0.500	0.676	0.762	0.762	0.548

The underlined values represent AUC>0.8; the blue-coloured values indicate the three best models; the grey-shaded attribute denotes the overall best model

In Table 14, CATA1_UK, CATA2_UK, CATA3_UK, CATA4_UK, CATA5_UK,and CATA5_UK represent models developed for the UK dataset using the CATA1, CATA2, CATA3, CATA4, CATA5, CATA7 groups, respectively.

Similar to the results for the Malaysian dataset, the results for the UK models showed that CATA1_UKand CATA2_UK were the worst models when compared to CATA3_UK, CATA4_UK, and CATA5_UK. Thus, again, suggesting that adding additional predictors, other than predictors of demographic, medical history, and medication received before admission categories, generally improved the performance of the prediction models. Beginning with the addition of predictors from the clinical investigation category (CATA3_UK), many algorithms achieved AUC > 0.8, such as BN, NB, LG, PART, ADT, FT, LMT, NBT, and RF. In fact, NB and ADT appeared to be the best classification algorithms for CATA3_UK (NB=0.832, ADT = 0.821), CATA4_UK (NB=0.848, ADT = 0.833), and CATA5_UK (NB=0.847, ADT = 0.841). Additionally, BN was also identified as the best classification algorithm for both CATA4_UK and CATA5_UK.

However, the performances displayed by DT, MLP, NBT, RT, and REPT dropped for CATA7_UK, which is similar to the scenario observed for the Malaysian dataset. Nonetheless, the best algorithms for CATA7_UK, which were NB and LMT, demonstrated improvement in their performances from AUC=0.847 to AUC=0.866, and AUC=0.830 to AUC=0.888, respectively.

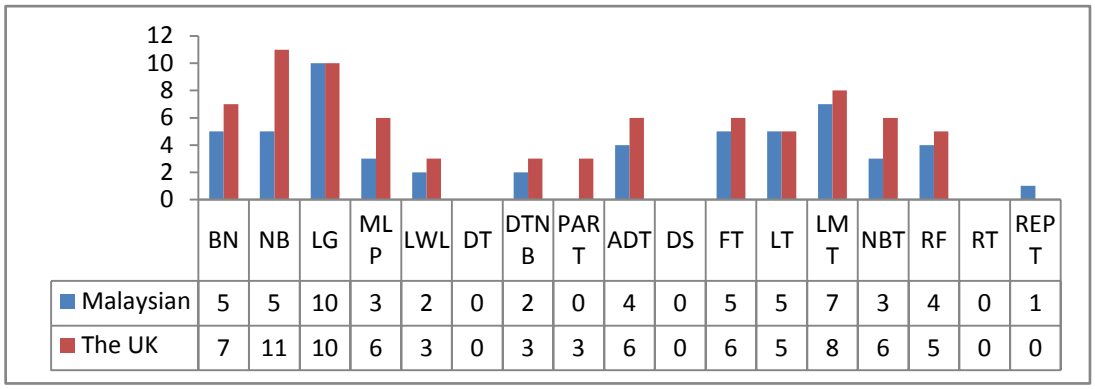
In conclusion, the predictors forCATA1 and CATA2 failed in generating convincing AUC results. In fact, the performances of the models began to improve upon inclusion of predictors from varying clinical categories. Furthermore, the results suggest that selecting essential and relevant predictors was more substantial than having a simple model with meaningless predictors. The inclusion of predictors solely from the demographic, medical history, and medication received before admission categories was proven to be insufficient in constructing good ACS models

Nevertheless, in CATA7, the number of predictors was reduced after filtering the predictors of each clinical category using a ML feature selection

method. Hence, the CATA7 models had been expected to enhance the performance of CATA5 via reduction of features. Although a drop was noted in the performances of the models developed on several algorithms, such as DT, MLP, NBT, RT, and REPT, most of the algorithms for both the Malaysian and UK datasets did show enhanced results, when compared to models with CATA5. Overall, the UK models exhibited more predictive scores than the Malaysian models.

5.2.7 Classification Algorithms on Feature Selection

This section presents the model performance of classification algorithms for the three main tasks in this chapter. A total of 16 input datasets were used to construct the three tasks using 17 ML algorithms. Note that models that achieved AUC>0.8 were assumed to be good. Thus, to reflect the best algorithms for model development, the frequency of each algorithm that achieved AUC>0.8 is depicted in Figure 12.



	BN	NB	LG	MLP	LWL	DT	DTNB	PART	ADT	DS	FT	LT	LMT	NBT	RF	RT	REPT
Malaysian	5	5	10	3	2	0	2	0	4	0	5	5	7	3	4	0	1
The UK	7	11	10	6	3	0	3	3	6	0	6	5	8	6	5	0	0

Figure 12: Frequency of classification algorithms producing good prediction models

Figure 12 presents the frequency of good models (AUC>0.8) by algorithm for the Malaysian and UK datasets. In total, 79 UK models attained an AUC> 0.8 compared to 56 Malaysian models. The three best algorithms for the UK models, based on the frequency of obtaining AUC>0.8, were NB, LG, and LMT. In addition, NB, LG, and LMT also emerged as the three best algorithms for the Malaysian models, despite the variability in frequency. Furthermore, DT, DS, RT, and REPT appeared to be unsuitable classification algorithms for both the Malaysian and UK datasets.

From another stance, the performance of ML algorithms was further scrutinized from the best models constructed from the three main tasks of this chapter. Table 15 summarizes three best models constructed for the three main tasks.

Table 15: The best models for evaluating ML feature selection method, evaluating predictors of existing ACS models and evaluating predictors of different clinical categories

Task	Methods	Malaysian	The UK	Comparisons
Evaluating ML feature selection method	**CFS**	LMT (0.802)	LMT (0.868)	
		LG (0.801)	LG (0.869)	
		NB (0.794)	NB (0.889)	
Evaluating predictors of existing ACS models	**GRACE**	NB (0.906)		0.826 (UK)
		LG (0.904)		0.827 (UK)
		FT (0.902)		0.827(UK)
	All_LR		NB (0.917)	0.845 (Malaysia)
			FT (0.902)	0.606(Malaysia)
			LG (0.874)	0.842(Malaysia)
Evaluating predictors of different category	**CATA7**	LMT (0.837)	LMT (0.888)	
		BN (0.850)		0.808 (UK)
		LG (0.837)		0.742 (UK)
			NB (0.866)	0.789 (Malaysia)
			ADT (0.835)	0.801 (Malaysia)

As illustrated in Table 15, in evaluating ML feature selection methods for both the Malaysian and UK datasets, the best models were constructed by applying the *CFS* method using the LMT, LG, and NB algorithms.

On evaluating predictors of existing ACS models, the best models developed for the UK dataset were found when adapting a combination of nine ACS models (All_LR), while the best models developed for the Malaysian dataset adapted predictors of GRACE models. Although the Malaysian and UK datasets revealed their best models from different sets of predictors, predictors of All_LR and GRACE were able to produce good models with the same ML algorithms, i.e., NB, LG and FT, with the exception of the FT algorithm on the Malaysian All_LR model.

Furthermore, on evaluating the effect of predictors of different clinical categories, the best models for the Malaysian and UK datasets were produced when predictors from the CATA7 group were used. The algorithms used to develop the Malaysian models were the LMT,BN, and LG algorithms, while the LMT, NB and ADT algorithms were used for the UK models.

5.3. Discussion

In this chapter, various ML feature selection methods were evaluated in order to produce better models. ML feature selection methods extract a set of predictors based on the patterns discovered (learning process of a machine) in a modelled dataset in order to simplified the model and gain good predictive power. The study also evaluated the potency of producing models by adapting predictors of existing ACS models using ML algorithms. Moreover, the effects of predictors from different clinical categories in constructing good models were also assessed.

Among the evaluated ML feature selection methods, the results suggest that *CFS* as the best method to identify the best set of predictors for both the Malaysian and UK datasets. Nevertheless, no clear conclusion could be made regarding the *subset* method being better than the *wrapper* method, mainly because the UK models displayed better discriminative ability when using the *wrapper* methods, as compared to the *FilterSubset* method. Meanwhile, the results for the Malaysian dataset support the findings from the study by Hall (2000), i.e. the *filter*(which includes the subset and filter method) methods performed better than the *wrapper* methods. Unfortunately, the results proved otherwise for the UK models. Nevertheless, this study has demonstrated that models developed based on sets of predictors selected by the ML algorithm possess competitive discriminative ability upon a number of classification algorithms. For example, among our best models, i.e., model developed based on the UK datasets, using sets of predictors identified by *CFS* method (CFS_UK), which were developed using NB, LG, MLP, FT, and LMT had better predictive power than TIMI, PURSUIT, GRACE, EMMACE, SRI, and C-ACS(Antman et al., 2000, Dorsch et al., 2001, Morrow et al., 2001, Huynh et

al., 2013, Boersma et al., 2000, Granger et al., 2003). This study has also demonstrated the potential of the ML feature selection method as an alternative way of selecting predictors for model development. In traditional statistical modelling, potential predictors are commonly pre-selected by considering clinical reasoning, reviewing the literature on existing models or known risk factors, and opinions from experts(Han et al., 2016). In addition, most of the predictors extracted from executing ML feature selection methods were also among the predictors of the existing ACS models, such as TIMI, GRACE, and PURSUIT, indicating that ML feature selection outcomes are consistent with the outcomes from traditional statistical modelling.

Overall, the results showed that most ML algorithms successfully achieved AUC>0.8 when adopting predictors from existing ACS models. Although both the Malaysian and UK datasets were inclined towards varied sets of predictors, the results indicate that the predictors of existing ACS models were indeed important predictors for the study's datasets, and, most probably, for ACS mortality models, in general. For Malaysian dataset, the best models was constructed adopting predictors from GRACE model (GRACE_MY) and for the UK dataset, the best models were constructed adopting predictors from a combination of predictors from 9 ACS models. Additionally, the models developed using ML algorithms displayed enhanced discriminatory ability when compared to those developed using traditional statistical methods. In fact, the best models constructed for the Malaysian and UK datasets(GRACE_MY and All_LR_UK)achieved competitively better predictive power in comparison to all the 11 reviewed ACS prediction models. As such, this study has established important supporting evidence of the use of ML algorithm in clinical prediction modelling.

Nevertheless, both the Malaysian and UK models adopting predictors of AMIS and Serbia models (AMIS_MY, AMIS_UK, SERBIA_MY and SERBIA_UK) failed to achieved better AUC value when compared one to one to the AMIS and Serbia model. Although the best models did achieve AUC > 0.8, the scores were still below the published AUCs of AMIS and Serbia models. AMIS model recorded an AUC of 0.875 on the AODE algorithm. AODE algorithm was not evaluated in this research as AODE

only accepts categorical attributes(thus suggesting a loss of information when dichotomizing continuous attributes, as mentioned previously). Additionally, the "pre-hospital cardiopulmonary resuscitation" attribute, an attribute in the AMIS model, was unavailable for the datasets employed in this study (Kurz et al., 2009).

On the other hand, the Serbia model recorded an AUC score of 0.91 using the ADT algorithm. The cohorts used for the Serbia model originated from ACS patients who had undergone PCI, which reflected cohorts of STEMI patients. Hence, the information in the predictors might not refer to the first entry for an event, which was applied as a criterion in this study's dataset. In fact, this study considered all ACS patients, including those diagnosed with NSTEMI and UA, in addition to STEMI. Furthermore, troponin was excluded as a predictor due to the quality issue for this predictor in our datasets. Biomarkers, such as troponin, are considered one of the essential predictors for ACS models(Khan et al., 2009, Granger et al., 2003). Other than that, the Serbia model applied cost sensitive learner to boost the predictive power of the ADT algorithm. These could be possible reasons for the notable differences in the AUC score obtained by the models developed in this study, in comparison to the Serbia model.

Furthermore, selecting potential predictors based on clinical category, which was also implemented by GUSTO-I(Steyerberg, 2009),highlighted that selecting predictors based solely on demographic and patients' history categories is insufficient in terms of producing a good prediction model. The performance of the model was enhanced after predictors from the clinical investigation, ECG, and baseline investigation categories were embedded in the model as predictors. This implies that most of the important predictors for an ACS prediction model come from a combination of the clinical investigation, ECG, and baseline investigation categories. Therefore, in order to construct a good ACS prediction model, the set of predictors must, at least, include predictors from the clinical investigation and ECG categories, in addition to the demographic and patient's medical history categories. Besides, the sets of predictors for the nine ACS models reviewed in this study were also stretched into a similar combination of clinical categories. Furthermore, by applying a feature selection method, i.e.

CFS, on these combinations of clinical categories, better AUC results were attained (CATA7). The model with *CFS* illustrated a simpler structure with less computational cost.

Nonetheless, predictors from the history of medication received by a patient category did not enhance the performance of the ACS prediction model. This was demonstrated by reconstructing models using theCATA3, CATA4, and CATA5 groups, but excluding the predictors from history of medication received by patient category. To be precise, better models were built without predictors from this category. Moreover, predictors under this category were rarely selected by most of the evaluated feature selection methods (refer Table 6). This outcome from the feature selection methods indicates additional supporting evidence that the predictors from the history of medication received by a patient category are not important for the ACS prediction model.

As concluded by Ali and Smith(2006),Tomar and Agarwal (2013),and Harper (2005),no specific classifier appeared to be best for all datasets. However, the findings obtained from this study suggest that a similar domain, i.e., ACS, with a similar target outcome and similar target ACS patients characteristics (e.g., first entry of ACS patients and all ACS types) may lead to the same best classification algorithms for the datasets. Perhaps the datasets with characteristics similar to the studied datasets would also attain the same best algorithms with which to construct a prediction model. Findings from the STATLOG studies, the largest algorithm comparison studies on a large number of different types of datasets, also concluded that the best algorithm to use mainly depends on the type of dataset being used(King et al., 1995).As such, this study concluded that LMT, BN, LG, NB, and ADT emerged as useful algorithms for the datasets, although slight variations were noted in the actual value of the predictive power (AUC). ADT was also found to be the best algorithm with which to construct ACS models for patients submitted for PCI (Sladojević et al., 2015). Nevertheless, DT, DS, RT, and REPT need to be added to the list algorithm found to be unsuitable in the previous chapter, which were VP, CR, Ridor, ZR, SVM, JRip, OneR, BFT, j48, SC, and KNN.

Finally, the three best models from each of the three main tasks of this chapter have been established. Although, the sample sizes for model development were reduced tremendously (approximately, on average, between 40-50% of the total baseline set for both datasets) due to missing values, we believe they are still sufficient to build reliable predictors. According to Mukherjee et al. (2003), for the ML classification problem, the minimum training size for model development in the treatment outcome problem is more than 50 samples. Besides, some well-developed and ACS models were also created from small sample sizes, such as the TIMI (n=1957) and EMMACE (n=3684) models, and and some of the latest ACS models, for instance, the C-ACS (n=4627), Serbia (n=2030), and MACE (n=2930) models. Furthermore, the best models established in this chapter have fewer predictors than the baseline models developed in Section 4.5.2. Since the dimensionality of the dataset has been reduced, the reduction in training sample size may not affect the outcome of the developed model.

5.4. Conclusion

ML feature selection has demonstrated its potential for identifying potential predictors for ACS prediction models and eventually constructing a competitive model. Comparing the *subset* and *wrapper* feature selection methods, the *CFS* method of *subset* feature selection emerged as the best method with which to determine the best set of predictors for the datasets. However, findings from the study are insufficient to conclude that the overall *subset* feature selection method is better than the *wrapper* feature selection method. In developing good ACS predictors, a combination of predictors should embed information, at least, from predictors in the demographic, patient medical history, clinical presentation, and ECG categories. In fact, predictors from only the demographic and patient medical history categories were proven to be insufficient for building competitive ACS prediction models. Furthermore, predictors from the medication received by patients category were found to be not important and had very little impact in terms of enhancing the performance of ACS models

Overall, this chapter has identified the most outstanding algorithms to be LMT, BN, LG, NB, and ADT for both datasets. And, the best sets of

predictors to construct ACS models from Malaysian dataset are : 1) age, heart rate, SBP, DBP, ECG Abnormalities - T-Wave inversion, and Lvef 2) age, heart rate, SBP, killip class, ACS symptoms before admission 3) age, history of premature CVD, history of heart failure, history of lung disease, history of renal failure, heart rate, SBP, DBP, ECG Abnormalities - T-Wave inversion, ECG Abnormalities - BBB, ECG Abnormalities - Non specific, ECG Abnormalities Location - Anterior Leads : V1 and V4, ECG Abnormalities Location - Right Ventricle : ST Elevation in Lead V4R, Low-density lipoprotein cholesterol(LDL-C), FBG, Lvef, Low molecular weight heparin (LMWH) taken, Angiotensin converting enzyme (ACE)inhibitors taken, diuretics taken, and anti-arrhythmic taken. As for the UK dataset, the best sets of predictors to construct ACS models are: 1) age, BB, SBP, cardiac arrest, and reinfarction2) age, gender, history of heart failure, on aspirin status, SBP, heart rate, cardiac arrest, ST-segment deviation of ECG 3) age, history of cerebrovascular disease, history of chronic renal failure, history of heart failure, diabetics, smoking status, aspirin status, BB, SBP, cardiac arrest, reinfarction, ECG, and tropinin assay.

Hence, the best models identified in this chapter were further validated and evaluated.

Chapter 6: Misclassification Analysis

This chapter presents the analyses performed for misclassification instances upon the 15 Malaysian and 15 UK models developed in Chapter 5. The objectives of this chapter are to explain the evaluation performed on the misclassified instances of these models to determine the reasons for misclassification. Furthermore, a prediction model developed to predict the misclassified instances will be presented.

6.1. Background

Performance measurements, such as accuracy, f-measures, precision, recall, and the AUC of classifiers, generally focus on the average or the overall performance of the constructed model. However, no information is given on misclassified instances and the reasons for misclassification. Misclassification instances generally occur due to 'bad' or 'noisy' data and/or attributes. 'Bad' data or 'noise,' on the other hand, is defined as the factor of confusion in building classification models, a factor which could negatively affect accuracy. Therefore, 'noisy' data or attributes must be reduced or eliminated so as to ensure the reliability of the model. As such, feature reduction is responsible for reducing 'noise' at the attribute level. At the instances level, 'noise' can be found mainly in the form of outliers (Seiffert et al., 2014) and overlapping classes(Smith, 2009, Stefanowski, 2013). In addition, the performance of a model may also be affected due to a skewed dataset (Visa and Ralescu, 2005), as well as the existence of small disjoints within a dataset (Jo and Japkowicz, 2004, Weiss, 2010). Hence, these are some factors that could contribute to misclassification in developing a model.

Moreover, upon understanding the reason behind the misclassified instances in a specific dataset, appropriate measures can be customised to handle a specific problem in a dataset. As noted by Smith and Martinez (2011), the identification of outliers is rather difficult as there are no generic definitions and characteristics of outliers. Thus, this study sought to

understand the reasons of misclassification in working with ACS datasets. Additionally, this study also explored the predictors derived from ACS datasets that contributed to misclassified instances while building the model. Predictors that contributed to misclassified instances for both the Malaysian and UK models were identified and used to develop models to predict misclassified instances for the datasets.

A number of studies have looked into identifying misclassified instances from a specific stance of misclassification. Specific to a breast cancer dataset, Thongkam et al. (2008)proposed a C-Support Vector Classification Filter (C-SVCF) to identify and remove the misclassified instances so as to improve model performance. The study used C-Support Vector Classification (C-SVC) with a radial basis kernel function to identify and eliminate outliers. In comparison to several of ensemble filter methods, such as AdaBoost, Bagging, and SVM ensembles, models using the C-SVCF method achieved better performance. In addition, Khoshgoftaar et al. (2004)presented a rule-based detection method using Boolean rules generated from the measurement data in detecting noisy instances. Earlier, Brodley and Friedl(1996) employed ensembles algorithms functioning as a filtering mechanism to eliminate misclassified instances of training data before actual model development. Furthermore, Smith (2009)investigated misclassified instances in a broader perspective. Their work analysed 190,000 instances from 64 datasets developed on nine different classification algorithms and concluded that five properties of instances were mostly likely to be misclassified. Moreover, class overlap appeared to be the main factor in instances of misclassification. This present research, on the other hand, focused on analysing misclassified instances only for ACS datasets, but from two varied populations. This is because understanding the clinical characteristics of misclassified instances could shed light on automation bias in an ACS clinical DSS. Eventually, this could further lead to the development of rule-based algorithms for DSS to reduce automation bias.

6.2. Method

6.2.1 Misclassification Analysis

The misclassification analysis was performed upon instances extracted from three Malaysian models and three UK models constructed on the five best algorithms, which were NB, BN, LG, ADT, and LMT. The three models employed were: 1) baseline models with missing values and outliers, 2) models developed based on a combination of predictors from 9 selected existing ACS models(All_LR), and 3) models developed based on predictors extracted from *CFS* method (CFS). These models were retrieved from Chapters 4 and 5. Note that the ALL_LR and CFS models were models developed based on datasets that are subsets of the baseline datasets.

Next, the identified misclassified instances were categorized as: 1) misclassified by > 3 algorithms, 2) misclassified by <3 algorithms, and 3) no misclassification by all the algorithms. Furthermore, the focus of the evaluation relied on the instances that were misclassified by at least half of the five algorithms. Therefore, from here on, in order to simplify the terms used, the notion "misclassified instances" or "misclassified cases" refers to instances misclassified by at least three algorithms. Furthermore, this study focused on the analysis of misclassified instances in minority classes, overlapping classes, outliers, and missing values. Hence, the frequency of misclassified instances against the five algorithms has been recorded.

A minority class refers to instances found in the smallest class of an imbalance or skewed dataset. In a classification task, a skewed dataset turns into an issue when the target class becomes the minority class. As such, a tendency to miscalculate the rate of accuracy is present as many classifiers only predict the majority class accurately, not the minority class.

On the other hand, an overlapping class reflects instances similar to those in another class. These instances were detected by using the simplest and widely used clustering method, the *K-Means* algorithm (using Euclidean distance)(Zhang et al., 2008b). The *K-Means* algorithm uses unsupervised learning, in which it partitions the datasets into k clusters by defining the cluster centre or mean (k-centroid) of each cluster. Thus, an initial k-centroid point is identified in the space of dataset objects. Next,

each object in the dataset is grouped based on the closeness of the object to a k-centroid point, which is calculated using Euclidean distance. After that, the positions of the k-centroids are recalculated, and the objects of the dataset are reassigned until the k-centroid points remain the same. In WEKA, the algorithm is called *SimpleKMeans*. Hence, to identify overlapped instances in the datasets, the instances were clustered based on the classes of the datasets.

Outliers in the studied context refer to values that were out of range of the specified range defined for an attribute.

The predictors of the models were also analysed to identify the patterns that indicate if a particular predictor could contribute to misclassified instances. The mean of the numerical values or the percentage of the categorical values of each predictor of misclassified instances were evaluated and compare against positive instances (died), negative instances (discharged), and overlapped misclassified instances. In addition, the missing values of the predictors of misclassified instances were also analysed.

6.2.2 Prediction Models for Misclassified Instances

The findings from analysing the predictors of misclassified instances turned into potential predictors that could have contributed to predict misclassified instances. These potential predictors were used to construct models so as to predict the misclassified instances. With that, the outcomes of the models were: 1) misclassified by > 3 algorithms, 2) misclassified by <3 algorithms, and 3) no misclassification by all algorithms. In fact, the predictors were obtained from the result of analysing predictors of misclassified instances. Furthermore, the models were built by using the five best algorithms identified in the previous chapters, which were BN, NB, LG, ADT, and LMT. However, ADT was dropped from the model development process as ADT could only accept problems with two classes. Next, the models were measured by using AUC and validated by using external datasets (the model developed using the Malaysian dataset was tested on the UK dataset and vice versa).

6.3. Results

6.3.1 Overall Misclassification Analysis

As illustrated in Figure 13, on average, the total misclassified instances were 6.5% for the Malaysian datasetand3% for the UK dataset. Hence, the UK models exhibited better overall discrimination power (results from Chapters 4 and 5) when compared to the Malaysian models. The following explains the variability in percentage of misclassified instances between these 2 datasets.

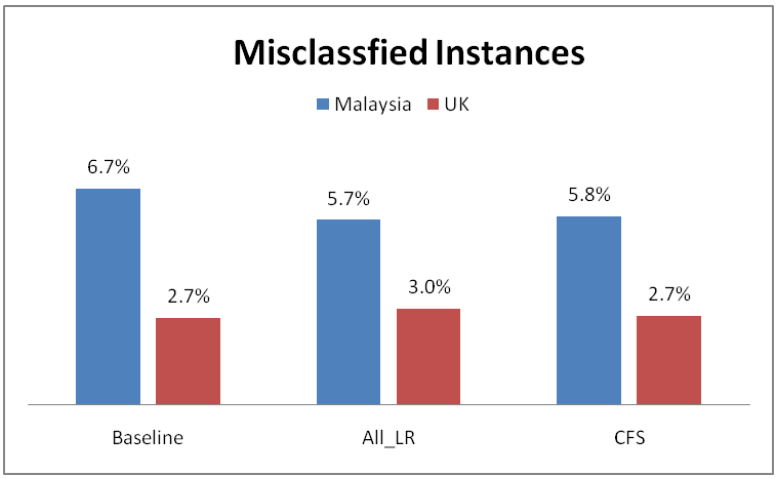


Figure 13: Percentages of Misclassified Instances - Malaysian Vs UK models

In Figure 13, a majority of the same misclassified instances in the baseline models had been observed across All_LR and CFS models. In addition, recall, ALL_LR, and CFS are models developed based on dataset that is subsets of the baseline dataset. Thus, for the Malaysian models, 81.3% of ALL_LR misclassified instances were also found to be misclassified in the baseline model, while 95.6% of ALL_LR misclassified instances were also misclassified in the CFS model. Furthermore, 81.7% of CFS misclassified instances were also misclassified in the baseline model. On the other hand, for the UK models, 69.4% of ALL_LR misclassified instances were misclassified in the baseline model, whereas 86.1% of the misclassified instances were also misclassified in the CFS model. Additionally, 64.1% of CFS misclassified instances were misclassified in the

baseline model, while 88.6% of *CFS* misclassified instances were also misclassified in the ALL_LR model.

6.3.2 Misclassification of Minority Classes

The results obtained for misclassification instances were further investigated by classes. The results indicate that the misclassified instances for both datasets leant towards the minority class (died). Furthermore, more than 85% of the misclassified instances lay on the minority class for both the Malaysian and UK models. Table 16 presents the details of misclassified instances by classes for the evaluated datasets.

Table 16 : Percentages of misclassified instances by classes

	Malaysia				UK			
	Died		Discharged		Died		Discharged	
	MI	Actual	MI	Actual	MI	Actual	MI	Actual
Baseline	394 (5.9%)	476 (7.1%)	53 (0.79%)	6197 (92.9%)	62 (2.3%)	122 (4.5%)	10 (0.4%)	2570 (95.5%)
All_LR	155 (5.6%)	162 (5.8%)	5 (0.2%)	2630 (94.2%)	41 (2.5%)	64 (3.9%)	8 (0.5%)	1583 (96.1%)
CFS	193 (5.6%)	202 (5.8%)	9 (0.3%)	3260 (94.2%)	33 (2.3%)	45 (3.2%)	6 (0.4%)	1379 (96.8%)

MI signifies misclassified instances

6.3.3 Misclassification on Overlapping

Overlapping instances were identified after executing the *K-Means* upon the datasets. Figure 14illustrates the overlapping percentage for each model. .

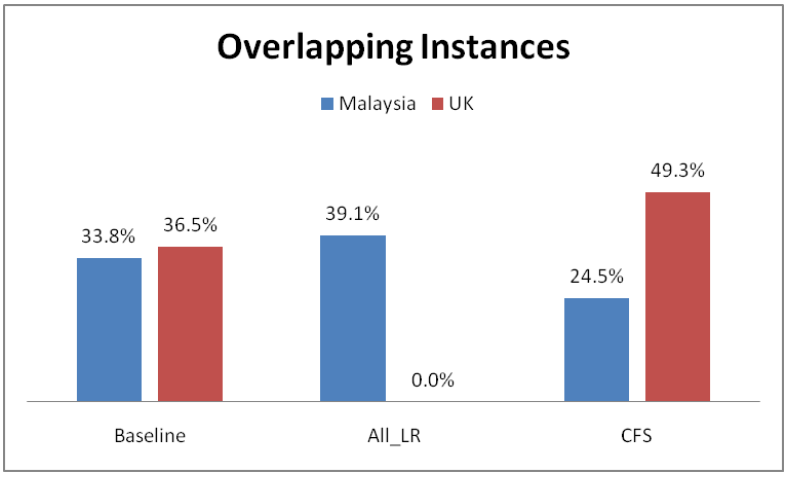


Figure 14: Percentage of Overlapping Instances

The percentages of overlapping instances for both the Malaysian and UK baseline models were similar as presented in Figure 14. In fact, changes in the overlapping percentage were observed for All_LR and CFS models, most likely due to the various combinations of predictors for All_LR and CFS, thus resulting in varied overlapping instances. For instance, 39.1% of the instances were overlapped in All_LR, while 24.5% of instances were found to overlap in CFS of Malaysian models. In addition, extreme variability in overlapped instances was observed in All_LR (0%) and CFS (49.3%) of the UK models. The detailed distribution of overlapping instances by classes is shown in Table 17. Furthermore, Table 18portrays the overlapping instances that were misclassified (O-Misclassified). From here on, to simplify, O-Misclassified is referred to overlapping instances that were misclassified.

Table 17: Overlapping instances by classes

	Malaysia			UK		
	Died	**Discharged**	**Total**	**Died**	**Discharged**	**Total**
Baseline	335 (5.0%)	1921 (28.8%)	2256 (33.8%)	78 (2.9%)	905 (33.6%)	983 (36.5%)
All_LR	89 (3%)	1002 (35.6.)	1091 (39.1%)	0	0	0
CFS	168 (5%)	681 (19.7)	849 (24.5%)	25 (1.8%)	677 (47.5%)	702 (49.3)

Table 18: Overlapping instances that were misclassifiedby classes

	Malaysian		The UK	
	Died	**Discharge**	**Died**	**Discharge**
Baseline	276 (4.1%)	15 (0.2%)	41 (1.5%)	4 (0.1%)
All_LR	86 (3.1%)	2 (0.1%)	NA	NA
CFS	160 (4.6%)	2 (0.1%)	15 (1.1%)	2 (0.1%)

Although the percentages of overlapping instances varied in each model (Figure 14), for the three Malaysian models and two UK models (dismissing ALL_LR for the UK dataset),the percentages of O-Misclassified in each class were similar, as illustrated in Table 18. For the Malaysian models, an average of 4% of O-Misclassified reflected cases that involved 'Died,' while 0.1% of O-Misclassified were on 'Discharged' cases. Meanwhile, for the UK models (excluding All_LR), the average percentage of O-Misclassified was 1.3% in 'Died' cases, whereas it was 0.1% for 'Discharged' cases.

From a different light, by comparing the percentages of overlapping instances that were misclassified (Table 18) with the total misclassified instances (Table 16), for the Malaysian models, more than half of the misclassified instances were found to be overlapped, i.e., Baseline - 65%, ALL_LR - 53%, and CFS - 80%. For the UK models, the number of misclassified instances that overlapped fluctuated between various models, for example, Baseline - 63%, ALL_LR - 0%, and CFS -30%. Nevertheless, in Malaysia and the UK models, more than 60% of overlapped instances in the minority class were misclassified. In conclusion, the results suggest that overlapping in the minority class contribute to misclassified instances. In fact, the overlapping instances in the minority class may indeed be the underlying cause of misclassification in the minority class.

6.3.4 Misclassification on Outliers

Outliers were only found in the baseline UK dataset. Table 19 presents results of outliers versus misclassified instances.

Table 19: Outlier instances that were misclassified and distributed by classes

	UK		
	Died	**Discharged**	**Total**
Overall Outliers	3 (0.1%)	30 (1.1%)	33 (1.2%)
Outliers -Misclassified	1 (0.0%)	0 (0%)	1 (0.0%)

The percentages of outliers were subtle (1.2% of the total dataset). Thus, the results suggestthat outliers did not contribute to misclassified instances, as only 3% of the total outliers were found to be misclassified.

6.3.5 Misclassification on Missing Values

Missing values were only available in the baseline models. Referring to Table 16, the comparison of percentages between misclassified instances in models with missing values (baseline) and models without missing values (All_LR and CFS) showed no obvious variability. Thus, from a higher level of observation, there is no obvious indication suggesting missing values as a main factor in misclassified instances.

6.3.6 Clinical Predictors on Misclassified Instances

This section illustrates the findings obtained from analysing the predictors of datasets against positive and negative instances, misclassified instances, and O-Misclassified. The missing values of each analysed group were also observed. Table 20 tabulates several selected key predictors of the models that highlight some essential patterns indicating the predictors that contributed to misclassified instances.

Table 20: Key predictors indicating potential in contributing to misclassified instances

Characteristics	Malaysia				The UK			
	Died	**Discharge**	**Misclassified**	**O-Misclassified**	**Died**	**Discharge**	**Misclassified**	**O-Misclassified**
	n= 476	*n= 6197*	*n=447*	*n=291*	*n= 122*	*n= 2570*	*n=72*	*n=45*
Age	65.8 (0.0%)	58.6 (0.0%)	**65.3** (0.0%)	**65.3** (0.0%)	79.9 (0.0%)	68.2 (0.0%)	**80.7** (0.0%)	**80.1** (0.0%)
Male	69.5% (0.0%)	76% (0.0%)	**67.6%** (0.0%)	**63.2%** (0.0%)	54.9% (0.0%)	64.4% (0.0%)	**52.8%** (0.0%)	73.3% (0.0%)
SBP	123.3 (4.8%)	140.2 (1.6%)	**125.7** (4.9%)	**127.2** (4.5%)	124.2 (29.5%)	150.6 (22.8%)	82.9 **(90.3%)**	96.6 **(86.7%)**
Heart rate	94.9 (4.0%)	82.8 (1.5%)	**94.2** (4.3%)	**97.2** **(5.2%)**	89.1 (31.1%)	83.9 (1.3%)	54.9 **(91.7%)**	70 **(88.9%)**
History of MI	13.9% (23.5%)	17% (20.8%)	**13.2%** (23.7%)	**14.8%** **(26.8%)**	27% (13.9%)	21.2% (9.8%)	**31.9%** (13.9%)	**31.1%** **(20.0%)**
History of heart failure	12.2% (18.9%)	6.1% (17.2%)	**11.6%** (19.2%)	**14.4%** (20.6%)	9% (23.0%)	5% (17.5%)	**9.7%** (20.8%)	**8.9%** **(28.9%)**
History of cerebrovascular	5.3% (21.6%)	3.2% (19.6%)	**5.1%** (21.5%)	**5.2%** (24.0%)	11.5% (23.0%)	7% (18.2%)	**12.5%** (20.8%)	**15.6%** **(28.9%)**
History of renal failure	10.7% (21.6%)	5.9% (19.4%)	**10.5%** (21.3%)	**12.4%** **(24.4%)**	9% (24.6%)	3.8% (17.9%)	**12.5%** (22.2%)	**8.9%** **(31.1%)**
History of hypertension	62.2% (17.0%)	60.6% (13.7%)	61.7% (17.9%)	66% **(18.2%)**	38.5% (18.0%)	41.4% (10.6%)	**37.5%** (13.9%)	26.7% **(22.2%)**
History of diabetics	45.2% (21.2%)	41.4% (16.8%)	**45.4%** (20.8%)	**52.6%** **(23.4%)**	15.6% (13.9%)	15% (8.9%)	**18%** (13.9%)	0% (20.0%)
History of lung disease	6.9% (19.3%)	2.8% (17.6%)	**6.7%** (19.0%)	**6.2%** **(21.6%)**	NA	NA	NA	NA
Aspirin taken	30% (10.3%)	32.1% (9.6%)	31.1% (9.6%)	34.4% **(12.0%)**	28.7% (5.7%)	21.2% (5.7%)	**33.3%** (4.2%)	**35.6%** (6.7%)
Current smoker	49.4% (7.4%)	56.7% (4.8%)	**48.8%** (6.7%)	41.2% (7.2%)	40.2% (19.7%)	57.4% (12.6%)	**37.5%** (19.4%)	**40%** **(26.7%)**
BB taken	19.5% (13.4%)	24.4% (11.8%)	**18.6%** (12.3%)	22% **(15.1%)**	23% (41.8%)	43% (28.7%)	**20.8%** (37.5%)	**17.8%** (42.2%)
Statin taken	26.3% (12.8%)	28.6% (11.5%)	27.3% (11.4%)	31.3% **(14.8%)**	36.9% (41.0%)	51.5% (28.5%)	**37.5%** (36.1%)	**40%** (40.0%)

ST elevation level 1	23.3% (0.0%)	19.5% (0.0%)	**22.6%** (0.0%)	10.7% (0.0%)	43.4% (3.3%)	35.2% (5.0%)	38.9% (4.2%)	37.7% (4.4%)
ST elevation level 2	39.5% (0.0%)	32.9% (0.0%)	**39.1%** (0.0%)	**37.8%** (0.0%)				
ST Depression	31.3% (0.0%)	26% (0.0%)	**31.1%** (0.0%)	**30.9%** (0.0%)	20.4% (3.3%)	15.8% (5.0%)	**25%** (4.2%)	**26.7%** (4.4%)
BBB	8.8% (0.0%)	4.4% (0.0%)	**7.8%** (0.0%)	**9.6%** (0.0%)	9.8% (3.3%)	4.6% (5.0%)	**8.3%** (4.2%)	**8.9%** (4.4%)

Values are number (%) or mean (standard deviation) [% of missing values]

The red-coloured values represent the value of predictors for misclassified or O-Misclassified cases that have the same pattern as the minority cases; the blue-coloured values denote the values of predictors for misclassified or O-Misclassified cases that have higher percentages of missing values compared to the minority cases.

From earlier findings, misclassified instances were found mostly in the minority class. Therefore, the distribution of predictors of misclassified instances was also geared towards the minority class. In addition, most of the O-Misclassified instances also seemed to follow the same pattern. Higher percentages of missing values were discovered in the O-Misclassified instances, as presented in Table 20 (refer to column O-Misclassified with the blue-coloured values). Although no obvious differences were found in the percentages of misclassified instances between the models with missing values and those without missing values, as concluded earlier, noticeably higher percentages of missing values in O-Misclassified hinted at the effect of missing values upon misclassified instances.

In addition, Table 21 lists all the predictors that indicated patterns suggesting a contribution to misclassified instances among the evaluated models. As such, they appeared to be potential predictors for constructing models meant to estimate misclassified instances. Twenty-four and 18 predictors were determined from the Malaysian and UK models, respectively. Furthermore, the blue-coloured predictors denote generic predictors for both the Malaysian and UK datasets.

Table 21: Potential predictors for predicting misclassified instances for the ACS dataset

Malaysian	The UK
Age*	Age*
Gender*	Gender*
SBP*	SBP*
Diastolic BP	Heart rate*
Heart rate*	History of MI*
History of MI*	History of heart failure*
History of cerebrovascular*	History of cerebrovascular*
History of heart failure*	History of renal failure*
History of diabetics*	History of hypertension
History of lung disease	History of diabetics*
Aspirin taken*	Aspirin taken*
Smoking status*	Current smoker*
BB taken*	BB taken*
Low molecular weight heparin taken	Statin taken
ACE Inhibitors taken	ECG - ST Depression*
Diuretic taken	ECG - BBB*
ECG - ST elevation Level 1	Cardiac arrest before admission
ECG - ST elevation Level 2	Reinfarction
ECG - ST Depression*	
ECG - T-Wave	
ECG - BBB*	
FBG	
Lvef	

The asterisked(*) predictors denote the generic predictors for both the Malaysian and UK datasets

6.3.7 Model to Predict Misclassified Instances

The potential predictors listed in Table 21 were first fed into the *CFS* feature selection method in WEKA. The set of predictors that were fed into WEKA were identified from: 1) all attributes listed in Table 21 for each of the Malaysian dataset (referred as MY_CFS) and the UK datasets (referred as UK_CFS),as well as 2) all common attributes of the Malaysian (referred as MY_C_CFS) and the UK datasets (referred as UK_C_CFS) listed in Table 21. The results obtained upon performing the *CFS* feature selection method are presented in Table 22.

Table 22: Sets of important features that predicted the misclassified instances

MY_CFS (8 attributes)	**MY_C_CFS (5 attributes)**	**UK_CFS (5 attributes)**	**UK_C_CFS (11 attributes)**
Age	Age	Age	Age
History of lung disease	History of heart failure	Gender	Gender
Heart rate	SBP	History of MI	History of MI
Diastolic BP	ECG - BBB	History of cerebrovascular	History of cerebrovascular
ECG - BBB	Death in hospital	History of renal failure	History of renal failure
FBG		History of heart failure	History of heart failure
Lvef		History of diabetics	History of diabetics
Death in hospital		Smoking status	Smoking status
		Aspirin taken	Aspirin taken
		BB taken	BB taken
		Statin taken	Death in hospital
		Reinfarction	
		ECG - ST Depression	
		ECG - BBB	
		Death in hospital	

Additionally, an additional predictor was embedded in each set of predictors identified, which was *overlapped*, indicating if the instance was indeed overlapped or otherwise. Furthermore, the models were constructed based on sets of predictors listed in Table 22.As a result, the related AUC scores of the models are presented in Table 23.

In Table 23, MY_CFS and UK_CFS represent models developed for the Malaysian and the UK dataset to predict misclassified instances, respectively. And, MY_C_CFS and MY_C_CFS represent models developed for the Malaysian and the UK dataset to predict misclassified instances using the common attributes of the two datasets, respectively. All the models were internally validated. On top of that, MY_C_CFS and UK_C_CFS were validated on external datasets.

Table 23: Performance of models to predict misclassified instance models

	Malaysia			*UK*		
	MY_CFS	**MY_C_CFS (internally validated)**	**MY_C_CFS (externally validated)**	**UK_CFS**	**UK_C_CFS (internally validated)**	**UK_C_CFS (externally validated)**
BN	0.94	0.907	0.709	0.853	0.845	0.793
NB	**0.953**	**0.911**	0.714	0.859	0.849	0.782
LG	0.939	0.888	**0.755**	**0.864**	**0.861**	0.804
LMT	0.94	0.886	0.742	0.863	**0.861**	**0.816**

The blue-coloured values denote the best models

Referring to Table 23, NB (AUC=0.953 and AUC=0.911) emerged as the best algorithm for predicting misclassified instances for the Malaysian dataset, while LG (AUC=0.864 and AUC=0.861) appeared to be the best algorithm for the UK dataset. Meanwhile, both MY_CFS and UK_CFS were revealed as the best models for predicting misclassified instances for each dataset. Nevertheless, when the models were built based on common predictors (MY_C_CFS and UK_C_CFS), their performances displayed a slight drop. Moreover, their performances continued to plunge when they were validated on external datasets. Overall, the models developed using the Malaysian dataset (MY_CFS and MY_C_CFS) obtained higher AUC scores compared to those developed using the UK dataset (UK_CFS and UK_C_CFS). Nonetheless, when tested on an external dataset, UK_C_CFS outperformed MY_C_CFS. Therefore, this study concluded that the best model for predicting misclassified instances for both the Malaysian and UK datasets was the UK_C_CFS, which was developed by using the LMT algorithm.

6.4. Discussion and Conclusion

This chapter concludes that misclassification in the ACS datasets was mainly due to an imbalanced dataset, in which most of the misclassified instances were derived from the minority class. In fact, this finding is in line with that obtained by Van Hulse et al.(2007), but differed from the results of Smith (2009),as his study discovered overlapping to be the main contributor to misclassified instances. Nevertheless, our study found that the overlapping instances in the minority class are indeed another major factor

in misclassified instances. Furthermore, missing values and outliers had very minimal impact upon misclassified instances. The impact of missing values on misclassified instances might be minimised by the strategies embedded in each algorithm to address missing values. Nonetheless, only a few key predictors were mostly affected by missing values, such as heart rate and SBP from the UK dataset.

In addition, potential predictors that predicted misclassified instances were determined by investigating the patterns of the predictors on misclassified instances. The outcome showed that each evaluated dataset had its own set of predictors to best predict misclassified instances. Moreover, the findings from the prior chapter suggested that the same classification algorithm performed the best and worst on a dataset from the same domain, i.e., ACS, with the same outcome and similar input characteristics. Therefore, the prediction model that predicted misclassified instances generally supported both datasets, and could, perhaps, support other ACS datasets as well that have characteristics similar to those of this study. However, further validation on the model has to be performed on other ACS datasets.

Furthermore, age, gender, history of MI, history of cerebrovascular, history of renal failure, history of heart failure, diabetes, smoking status, aspirin taken, BB taken, and death in hospital functioned as the generic set of predictors that estimated misclassified instances for both the Malaysian and UK datasets. Therefore, a promising prediction model that predicted misclassified instances was generated using the UK dataset with these predictors. The model achieved an AUC = 0.861 when validated on the UK dataset, while it achieved an AUC=0.816 upon validation on an external dataset (Malaysian dataset). Thus, the models may add input to addressing the automation bias issue in the context of ACS prediction modelling. In addition, a rule-based algorithm may be developed based on the model and findings, primarily to decrease automation biasness in a DSS of ACS domain.

The findings from this particular chapter highlight the major contribution that could affect the performance of ACS prediction models. In

fact, the same issue may affect dataset with similar characteristics to those of study. Hence, further studies concerning ACS prediction modelling should be targeted on resolving issues related to imbalanced datasets and overlapping of minority classes.

The following chapter depicts several strategies that could be implemented to address issues related to imbalanced datasets and missing values found among key predictors.

Chapter 7: Model Optimization

This chapter presents the proposed methods for handling imbalanced datasets and missing values so as to enhance the performance of the models.

7.1. Background

The results obtained from the analysis of misclassification instances advocated that the main contributor of misclassified instances within the datasets was an imbalanced dataset. In fact, many cases concerning misclassified instances involved the minority class. Besides, this finding was supported when a "RemoveMisclassified" function in WEKA was performed on the datasets. "RemoveMisclassified" refers to a function found in WEKA that eliminates expected misclassified instances. Thus, when the function was executed on both datasets, all instances of minority instances were discarded. In addition, cases of overlap were also found to have a notable contribution to misclassified instances among the datasets. Further, the findings obtained by Denil and Trappenberg (2010) and Lopez et. al.(2013) implied that overlapped classes do hinder the performance of a classifier. Therefore, based on these two factors, this study proposed a new strategy to address issues related to imbalance datasets using the *undersampling* method(Liu et al., 2009).

Furthermore, the results from misclassification analysis also showed that missing values had a very minimal impact on misclassified instances. But, some of the key predictors showed a strong effect on misclassified instances. Therefore, in handling the missing values of the datasets, this study proposed a method called the *mean-clustering-imputation* method in dealing with the missing data.

7.2. Method

7.2.1 Overlapped-Undersampling Method

In the *undersampling* method, the majority class is resampled to decrease the biasness of the majority class. The simplest *undersampling* method refers to the *random undersampling* method, in which instances in the majority class are removed until a fair distribution is achieved.

This study proposed a strategy in which the instances to be deleted from the majority class were the overlapped instances of the majority class. The method is referred as the *Overlapped-undersampling* method. *SimpleKMeans* was applied to determine the overlapped instances. All of the overlapped instances from the majority class were then discarded from the training set. After that, the AUC results of the new strategy were compared with the following approaches so as to address the issue of an imbalanced dataset:

1) No sampling method - the dataset was used as it was
2) *Random undersampling* method - the instances in the majority class were discarded in a random manner(Yap et al., 2014). The ratio of majority and minority classes adhered to the ratio in *Overlapped-undersampling.*
3) *Boosting* - An ensemble method using similar classifier algorithms. The model is iteratively built based on the weight of each instance for each iteration, with initially equal weight for all instances. In each iteration, each instance is assigned a greater weight for a misclassified instance and a lower weight for a correctly classified instance. This method is also called the *AdaBoost* in WEKA(Freund and Schapire, 1996). The ADT is another Boosting method that uses DT based on the classifier found in WEKA.
4) *Bagging* - An ensemble method using the same classifier. This model is developed via random sampling replacement in each iteration (Breiman, 1996).

5) *Voting* - An ensemble method using different classifier algorithms(Kittler et al., 1998). The classifiers employed in the method were BN, NB, LG, ADT, and LMT.

6) *Random Forest* (RF) - An ensemble method using tree classifier algorithms. The model is developed by inducing bootstrap samples with random feature selection in the tree induction process (Breiman, 2001).

All of the above approaches (except for *Voting* and RF) were run on the five best algorithms determined in this study, which were BN, NB, LG, ADT, and LMT.

7.2.2 Mean-Clustering-Imputation Method

One way of handling missing data in predictive modelling is to use the imputation method. The imputation method is a process of substituting a missing value with a value. The value can be decided either by identifying a globally constant or mean value or by identifying the most probable value. The constant or mean value assumes the all missing values are of the same value, and this may lead to distortions in the data's distribution. On the other hand, the proposed method, i.e. *mean-clustering-imputation*, proposes that the mean value (for numerical attributes) or most frequent value (for categorical attributes) is calculated by clustered samples, instead of a single value. This means that the training sample is first clustered using *Simple EM (expectation maximisation)*, while ignoring class labels. Then, for each cluster, the mean (for numerical attributes) or the most frequently occurring value (for categorical attributes)are calculated as the imputation values for missing data. Hence, each cluster has its own mean or most frequently occurring values to be imputed. By grouping the instances of similar groups, the most probable means or most frequently occurring values can be acquired. The approach of the method is slightly different from the latest Hruschka et al.(2004) approach. In their study, the clustering was done using a K-Means algorithm, then their method was applied to complete instances according to class labels, and the imputed values were calculated by finding the means of the corresponding attribute values of similar complete instances.

Thus, all the missing values were filled in both datasets with the imputation values calculated using *mean-clustering-imputation* method. After that, the AUC results of the proposed imputation method were compared with the following approaches so as to address the issue of missing values:

1) *Single-imputation* used mean values for continuous attributes and most frequently occurring values for categorical attributes
2) removed instances with missing values
3) allowed the algorithm to handle the missing values. In WEKA, each algorithm has its own way of handling missing values. The strategy for handling the missing values is embedded in the algorithm.
    - BN - uses *ReplaceMIssingValuesFilter,* which replaces the missing values with the mean (for numerical attributes) or the most frequent value (for categorical attributes)(Bouckaert, 2008).
    - NB - ignores the missing attributes(weka.classifiers.bayes.NaiveBayes) (John and Langley, 1995)
    - LG - uses a *ReplaceMIssingValuesFilter,* which replaces the missing values with the mean (for numerical attributes). All the categorical attributes are transformed into binary attributes using a *NominalToBinaryFilter*(weka.classifiers.functions.Logistic)(Le Cessie and Van Houwelingen, 1992)
    - ADT - the missing values are not propagated down the subtrees(Freund and Mason, 1999).
    - LMT - replaces the missing values with the mean (for numerical attributes) or the most frequent value (for categorical attributes)(Landwehr et al., 2005)

The above approaches were run on the five best algorithms determined in this study, which were BN, NB, LG, ADT, and LMT.

7.3. Results

7.3.1 Overlapped-Undersampling Method

The results that were obtained after applying the approaches in handling imbalanced datasets are tabulated in Table 24.

Table 24:Comparison of model performance with varied approaches in handling imbalanced datasets

	Malaysia					The UK				
	BN	**NB**	**LG**	**ADT**	**LMT**	**BN**	**NB**	**LG**	**ADT**	**LMT**
CATA7	**0.850**	**0.789**	**0.837**	**0.801**	**0.837**	**0.808**	**0.866**	**0.742**	**0.835**	**0.888**
CATA7_RM_Ovlp	0.830	0.793	0.811	0.707	0.830	0.808	0.863	0.735	0.846	0.813
CATA7_RandUdrSmp	0.850	0.786	0.827	0.803	0.833	0.812	0.872	0.754	0.820	0.879
CATA7_ADA_Boost	0.539	0.714	0.728	0.751	0.734	0.716	0.775	0.702	0.737	0.629
CATA7_Bagging	0.822	0.789	0.832	0.837	0.769	0.831	0.866	0.776	0.842	0.831
CATA7_RF			0.803					0.693		
CATA7_Voting			0.830					0.854		
GRACE	**0.844**	**0.906**	**0.904**	**0.789**	**0.868**	**0.818**	**0.826**	**0.827**	**0.786**	**0.83**
GRACE_RM_Ovlp	0.692	0.709	0.694	0.681	0.693	0.743	0.766	0.827	0.758	0.72
GRACE_RandUdrSmp	0.808	0.826	0.826	0.799	0.824	0.843	0.907	0.900	0.789	0.895
GRACE_ADA_Boost	0.766	0.695	0.713	0.777	0.699	0.713	0.867	0.870	0.842	0.804
GRACE_Bagging	0.828	0.825	0.825	0.826	0.810	0.890	0.903	0.902	0.886	0.893
GRACE_RF			0.828					0.782		
GRACE_Voting			0.833					0.879		
ALL_LR	**0.843**	**0.845**	**0.842**	**0.807**	**0.861**	**0.872**	**0.917**	**0.874**	**0.799**	**0.639**
ALL_LR_RM_Ovlp	0.696	0.726	0.710	0.678	0.714			NA		
ALL_LR_RandUdrSmp	0.764	0.799	0.793	0.763	0.789			NA		
ALL_LR_ADA_Boost	0.742	0.671	0.721	0.792	0.720	0.743	0.880	0.808	0.855	0.829
ALL_LR_Bagging	0.787	0.800	0.804	0.792	0.705	0.908	0.917	0.887	0.882	0.881
ALL_LR_RF			0.821					0.840		
ALL_LR_Voting			0.794					0.883		
CFS	**0.762**	**0.794**	**0.801**	**0.773**	**0.802**	**0.793**	**0.889**	**0.869**	**0.794**	**0.868**
CFS_RM_Ovlp	0.679	0.702	0.703	0.698	0.484	0.681	0.726	0.781	0.635	0.607
CFS_RandUdrSmp	0.766	0.795	0.801	0.763	0.500	0.834	0.890	0.872	0.801	0.848
CFS_ADA_Boost	0.698	0.705	0.732	0.769	0.681	0.630	0.784	0.800	0.806	0.694
CFS_Bagging	0.784	0.796	0.801	0.785	0.737	0.846	0.887	0.870	0.848	0.871
CFS_RF			0.758					0.713		
CFS_Voting			0.798					0.861		

The red-shaded rows are models that were developed without any optimization strategy; the grey-shaded rows denote the results of the proposed method; the red-coloured values indicate that the models with an optimization strategy outperformed the models with anon-optimization strategy.

In Table 24, the extension name attached to each model i.e. RM_Ovlp, RandUdrSmp, ADA_Boost, Bagging, RF and Voting represent the *Overlapped-undersampling, Random undersampling, Boosting*, *Bagging, Random Forest* and Voting approaches, respectively, in handling imbalanced dataset.

The results of Table 24 show that no improvement was established upon implementing the new strategy (rows shaded in grey)to address an imbalanced dataset. In fact, in most cases, they attained the lowest AUC score when compare to other approaches.

To be precise, all the imbalanced optimisation approaches were found to be inappropriate for the Malaysian dataset. In fact, the Malaysian dataset was better enhanced when no optimisation strategy was employed (rows shaded in red). Unlike the Malaysian dataset, the UK dataset demonstrated improvement when *Bagging* and *Random undersampling* approaches were applied.

7.3.2 Mean-Clustering-Imputation Method

Table 25 compares the AUC results of the proposed imputation method, i.e., *mean-clustering-imputation* , with *single-imputation* method*,* no missing values in the datasets, and using the method embedded in a algorithm.

Table 25: Comparison of model performance with varied approaches in handling missing values

	Malaysia					UK				
	BN	**NB**	**LG**	**ADT**	**LMT**	**BN**	**NB**	**LG**	**ADT**	**LMT**
	Missing Numerical/Categorical attributes - 1/2					***Missing Numerical/Categorical attributes - 2/1***				
GRACE_MEAN_CLSTR_IM	0.794	0.796*	0.798	0.783*	0.797	0.803	0.830*	0.830*	0.806*	**0.828**
GRACE_MEAN_IM	0.793	0.795	0.797	0.783	0.797	0.803	0.833	**0.831**	**0.816**	0.827
GRACE_NO_MISSING	**0.844**	**0.906**	**0.904**	0.789	**0.868**	**0.818**	**0.826**	0.827	0.786	0.826
GRACE_ALGRTHM	0.777	0.805	0.797	**0.794**	0.797	0.791	0.834	0.830	0.808	0.827
	Missing Numerical/Categorical attributes - 4/3					***Missing Numerical/Categorical attributes - 2/4***				
All_LR_MEAN_CLSTR_IM	0.758	0.730	0.778	0.765*	0.779	0.813	0.826*	0.825	0.803	**0.826**
All_LR_MEAN_IM	**0.754**	0.767	0.776	0.764	0.776	0.804	0.825	0.824	0.806	0.825
All_LR_NO_MISSING	**0.769**	**0.800**	**0.798**	0.734	**0.793**	**0.872**	**0.917**	**0.874**	0.799	0.639
All_LR_ALGRTHM	0.746	0.776	0.776	**0.782**	0.776	0.794	0.829	0.824	**0.808**	0.824
	Missing Numerical/Categorical attributes - 4/0					***Missing Numerical/Categorical attributes - 1/3***				
CFS_MEAN_CLSTR_IM	**0.771**	0.768*	0.774	0.765	0.774	0.787	0.815*	0.797	0.808*	0.825
CFS_MEAN_IM	0.761	0.503	0.742	0.736	0.742	0.785	0.813	0.822	0.805	0.823
CFS_NO_MISSING	0.762	**0.794**	**0.801**	**0.773**	**0.802**	**0.793**	**0.889**	**0.869**	0.794	**0.868**
CFS_ALGRTHM	0.751	0.777	0.774	0.760	0.772	0.785	0.828	0.822	**0.825**	0.822
	Missing Numerical/Categorical attributes - 6/8					***Missing Numerical/Categorical attributes - 1/11***				
CATA7_MEAN_CLSTR_IM	0.830	0.788*	0.804	0.795	0.801	0.784	0.812	0.780	0.752	0.784
CATA7_MEAN_IM	0.800	0.785	0.801	0.740	0.790	**0.816**	0.821	**0.806**	0.802	0.810
CATA7_NO_MISSING	**0.850**	0.789	**0.837**	**0.801**	**0.837**	0.808	**0.866**	0.742	**0.835**	**0.888**
CATA7_ALGRTHM	0.776	**0.803**	0.798	0.773	0.785	0.806	0.834	0.805	0.83	0.81

The grey-shaded rows denote the results of the proposed method; the blue-coloured values indicate the best models for a particular algorithm; the underlined values indicate that the proposed method is better than the MEAN_MI and ALGRTHM methods; the asterisked (*) values indicate that the proposed method is better than MEAN_MI method

In Table 24, the extension name attached to each model i.e. MEAN_CLSTR_IM, MEAN_MI, NO_MISSING and ALGRTHM represent the *mean-clustering-imputation* method, *single-imputation* method*,* no missing values in the datasets, and using the method embedded in a algorithm, respectively, in handling missing values.

Overall, the results in Table 25show that the best models were achieved when all instances with missing values were removed from the training sets (NO_MISSING). This scenario was observed in both the Malaysian and UK datasets.

Specifically for the Malaysian datasets, the results highlighted that the proposed methods were generally better than the MEAN_IM and ALGRTHM methods, specifically on the BN, LG, and LMT algorithms. In addition, in most of the other algorithms, the proposed methods were better than the MEAN_IM method.

On the other hand, for the UK dataset, the proposed methods produced the worst models on the CATA7 dataset. However, for the other UK data, the results demonstrate almost similar results to the Malaysian data, in which the BN, LG, and LMT algorithms, incorporating the proposed *imputation* method, built better models compared to models built with the MEAN_IM and ALGRTHM methods. And, in most of the other algorithms, the proposed method produced better models than models developed using the MEAN_IM method.

Even though models using the MEAN_CLSTR_IM method built better models in most cases compared to the models applying the MEAN_IM and ALGRTHM methods, the improvements of these models were not vivid (i.e. the AUC score was the same or very minimally increased by 0.001 or 0.002).This scenario might be due to the fact that the imputed values for missing values of categorical attributes in each cluster were the same, i.e. the frequent value of the attribute. For example, the most frequent value of a categorical attribute in Cluster 1 was 'No.' The value was noted to be the same in all other clusters with the same categorical attribute. Hence, all the categorical attributes with missing values in Malaysian dataset were had this scenario. Similar scenarios were also observed in all the UK datasets

except for the CATA7 datasets. To the contrary, in CATA7, there were three categorical attributes with missing values that had various imputed values. Surprisingly, applying the MEAN_CLSTR_IM method to theCATA7 dataset did not help improve the classification performance. In fact, the MEAN_IM and ALGRTHM methods produced considerably better models on the CATA7 dataset.

Nevertheless, better models were built with the proposed *imputation* method when the number of numerical attributes with missing values exceeded two in a dataset. This scenario was observed in CFS and CATA7for the Malaysian dataset. However, as noted in Table 27, none of the UK datasets had more than two numerical attributes with missing values.

7.4. Discussion and Conclusion

From the findings obtained regarding the effect of overlapping instances upon misclassified instances, this study suggested a new strategy to overcome the issue of imbalanced datasets via the *undersampling* method. This suggested strategy removed overlapping instances found in the majority class from the training set. Nonetheless, the strategy failed to produce satisfactory results. This was probably due to the removal of important information from the training data. The deleted overlapped instances might have actually contained some vital information for training a model. Furthermore, the Malaysian dataset performed better when no approach was taken to tackle an imbalanced dataset. Perhaps, the size of the training set for the Malaysian dataset could be the reason for this scenario. As according to Japkowicz et al. (2002), when a sample is large enough to represent sub-clusters in each class, an imbalanced dataset does not hinder the performance of a classifier. The sample size of the Malaysian dataset was obviously larger than that of the UK dataset. On the other hand, the UK dataset displayed better results when *Bagging* and *Random undersampling* approaches were applied. Therefore, UK models with applications of *Bagging* and *Random undersampling* approaches were further validated, with the details presented in the next chapter.

Moreover, in terms of handling missing values, this study proposed a method for establishing the imputed value for missing data. The method was named the *mean-clustering-imputation* method. In the method, training sample were first clustered. The imputed value was established by calculating the mean (for numerical attributes) or the most frequent value (for categorical attributes) of each cluster. The results demonstrated that removing instances with missing values resulted in the best models produced for both the Malaysian and UK datasets. However, removing instances with missing values could result in reducing the sample size enough to affect the reliability of the model. Thus, in this study, we applied feature selection before removing the instances with missing values. By reducing the number of predictors for model development (feature selection), the number of instances with missing values were eventually reduced, and, thus, an appreciable amount of observations remained for the training samples. Nevertheless, this approach is only applicable if a dataset is large enough to maintain a reasonable number of observations for the training samples after removing the instances with missing values.

Better models were constructed with the proposed imputed method compared to the *single-imputation* method and methods embedded in an algorithm, specifically when the models were developed on BN, LG, and LMT algorithms. In fact, the proposed method built notably satisfactory models when the number of numerical attributes with missing values was greater than two. Otherwise, the performance of models using the proposed imputed method were about the same or slightly better than models developed using the *single-imputation* method and methods embedded in an algorithm. However, the performance of the models using the proposed imputed method showed no improvement when there were more missing values in categorical attributes as opposed to numerical attributes.

Chapter 8: Model Validation

This chapter presents the model validation process of the best selected models from the previous chapter. The objective of the chapter is to further validate the models using internal and external datasets.

8.1. Method

The models were tested for internal and external validation. Internal validation involves testing the models using similar underlying populations, whereas external validation denotes testing the models on other populations. The final stage is to present the calibration so as to evaluate the performances of the predictions versus the actual outcome of the best models. Lastly, the overall calibration performances of the best models are measured by using the BS, while the visual agreement of the actual outcomes and predictions are presented on calibration plots.

8.1.1 Internal Validation

The best models identified in Chapter 5, which were CATA7, ALL_LR, GRACE, and CFS, were validated against the testing dataset using the five best algorithms: BN, NB, LG, ADT, and LMT. The testing set was reserved earlier during the pre-processing phase, as elaborated in Chapter 4. No exclusions were made on the testing dataset, except for incomplete cases with missing values. Initially, a total of 3,178 testing observations were assigned to testing for the Malaysian dataset, whereas 1,283 observations were assigned to testing for the UK dataset. Both of these numbers were assigned prior to discarding incomplete cases. Table 26 tabulates the total testing samples for each model employed for internal validation after the exclusion of incomplete cases.

Table 26 : Summary of testing samples for internal validation

Model	Malaysia (n)	UK (n)
CATA7	740	536
ALL_LR	1353	780
GRACE	2316	925
CFS	1656	658

Table 26 illustrated that the testing size for almost all models has been reduced to as low as 50% of the originally reserved observations for the testing set. The largest loss of testing observations was observed on the Malaysian CATA7 model, which contained only 740 observations. Despite the reduction of the testing set, the sample is still reasonable for validation. Results from Beleites et al.'s(2013) study suggested that a minimum of 75-100 samples is required to achieve reasonable precision in validating an ML classification model. Thus, the testing samples used to validate our models were sufficient.

As the UK models exhibited improvements upon applying the *random undersampling* approach, all models developed with the *random undersampling* approach were also validated.

The best models (CATA7, ALL_LR, GRACE, and CFS) was also prepared to be validated on the best generic model. The generic model is a model that is suitable for both the Malaysian and UK datasets. The predictors of the generic model should be common to both datasets. Thus, the best predictors were adjusted to only consider common predictors of the two datasets. As such, CATA7_CMM, ALL_LR_CMM, GRACE_CMM, and CFS_CMM were referred to as the generic models of CATA7, ALL_LR, GRACE, and CFS, respectively. The same testing samples (as in Table 26) were also used to validate these generic models.

8.1.2 External Validation

In order to validate a model, the testing set must have similar predictors as those found in the derived model. For instance, if the model is built with four predictors: 1) age,2) heart rate, 3) SBP, and 4) height, similar predictors must also exist in the testing set. Hence, only generic models are applicable for external validation. Among the generic models (CATA7_CMM, ALL_LR_CMM, GRACE_CMM, and CFS_CMM), only the best models identified from internal validation were further validated on an external dataset.

8.2. Results

8.2.1 Internal Validation

Tables 27,28, and 29present the results of internal validation. As such, Table 27 illustrates the results of the original best models, while Table 28 shows the results of the best models using the *undersampling* method, and Table 29 portrays the results of the generic models. In the validation process, the study considered a model with an AUC score of 0.75 and above as a good model.

Table 27: Results of internal validation of the best models

	Malaysia					UK				
	BN	NB	LG	ADT	LMT	BN	NB	LG	ADT	LMT
CATA7	0.798	0.756	0.724	0.765	0.739	0.810	0.772	0.563	0.732	0.679
ALL_LR	0.781	0.768	0.770	0.733	0.767	0.816	0.836	0.792	0.795	0.639
GRACE	0.824	**0.827	0.822	0.797	0.822	0.811	0.828	0.773	0.790	**0.847
CFS	0.753	0.755	0.762	0.733	0.760	0.770	0.781	0.685	0.747	0.786

The coloured values indicate the models with AUC>0.75; the grey-shaded values denote the best model for each set of models; the double asterisked (**) values represent the best model for the Malaysian and UK datasets

The results of Table 27showthat the best models for Malaysian and the UK datasets were models adopting predictors from GRACE model (GRACE). The best Malaysian model was developed by using NB, whereas the best UK model was constructed on the LMT algorithm. Nonetheless, NB and LMT emerged as the two best algorithms for constructing an ACS prediction model for both the Malaysian and UK datasets. The GRACE of

Malaysian model built using LMT algorithm had a slightly lower AUC score of 0.822in comparison to an AUC of 0.827 for NB. Similarly, for the GRACE of UK model, the NB algorithm also displayed the ability to generate a considerably good prediction model with an AUC of 0.828.

Table 28: Results of internal validation - Models with *random undersampling* method (UK dataset)

	UK				
	BN	**NB**	**LG**	**ADT**	**LMT**
CATA7	0.810	0.772	0.563	0.732	0.679
CATA7 _RandUdrSmp	0.799	0.766	0.526	**0.733**	**0.727**
ALL_LR	0.816	0.836	0.792	0.795	0.639
ALL_LR _RandUdrSmp			NA		
GRACE	0.811	0.828	0.773	0.790	**0.847**
GRACE_ RandUdrSmp	**0.812**	0.828	**0.774**	**0.792**	0.840
CFS	0.770	0.781	0.685	0.747	0.786
CFS_ RandUdrSmp	**0.783**	0.781	**0.686**	0.746	0.777

The underlined values indicate that the models with the *random undersampling* method are better than the models without the *random undersampling* method; the blue-coloured value represents the best model

Furthermore, it is important to note here that only the UK models had demonstrated improvements when the *random undersampling* approach was implemented. Although the models with the *random undersampling* approach seemed to have enhanced performance during development, these performances were generally at par with those models without the *random undersampling* approach when tested upon testing datasets. Table 28compares the results of internal validation of the models for both with and without the application of the *random undersampling* approach. CATA7_RandUdrSmp, All_LR_RandUdrSmp, GRACE_RandUdrSmp, and CFS_RandUdrSmp represent CATA7, All_LR, GRACE, and CFS models with *random undersampling* approach.

Out of the tested 15 models, only seven models appeared to have a higher AUC in comparison to models without the *random undersampling* approach. Furthermore, only CATA7_RandUdrSmpwith the LMT algorithm and CFS_RandUdrSmp with the BN algorithm displayed improvements upon implementation of the *random undersampling* approach. Otherwise,

improvement by 0.001 were noted for GRACE__RandUdrSmp on BN, CATA7_RandUdrSmp on ADT, and CFS_RandUdrSmp on LG. In addition, eight models exhibited lower performance than those without the *random under sampling* approach. Therefore, this study concluded that imposing a *random undersampling* approach had no notable contribution towards enhancing overall model performance.

Table 29: Results of internal validation for generic models

	Malaysia					UK				
	BN	**NB**	**LG**	**ADT**	**LMT**	**BN**	**NB**	**LG**	**ADT**	**LMT**
CATA7_CMM	**0.764**	0.756	0.749	0.710	0.500	**0.744**	0.726	0.741	0.645	0.500
ALL_LR_CMM	0.761	0.759	**0.765**	0.729	0.500	0.724	0.753	**0.767**	0.636	0.500
GRACE_CMM	0.742	0.771	0.754	0.726	****0.773**	0.705	0.743	**0.747**	0.711	0.500
CFS_CMM	0.713	**0.751**	0.748	0.704	0.500	0.761	0.774	**** 0.779**	0.756	0.500

The underlined values indicate the models with AUC>0.75; the blue-coloured values denote the best model for each set of models; the double asterisked (**) values represent the best model for the Malaysian and UK datasets

As expected, the performance of the generic models as represented in Table 29was lower than that of the models exclusively developed for each individual dataset. Nonetheless, the best generic model developed based on the Malaysian dataset was the GRACE_CMM using the LMT algorithm (AUC=0.773), whereas the best generic model built based on the UK dataset was the CFS_CMM using the LG algorithm (AUC= 0.779).

8.2.2 External Validation

External validation was performed only on Malaysian generic model of GRACE_CMM and the UK generic model of CFS_CMM, as they appeared to be the best generic models discovered from internal validation. The Malaysian generic model was validated on the testing set of the UK dataset (n= 925), whereas the UK generic model was validated on the testing set of the Malaysian dataset (n=1424). Table 30 presents the result of external validation of the generic models. As in Table 30, the results of external validation are compared with the results of internal validation.

Table 30: Results of external validation

	Malaysia		UK	
	GRACE_CMM	GRACE_CMM_Ext	CFS_CMM	CFS_CMM_Ext
BN	0.742	0.665	0.761	0.607
NB	0.771	**0.708**	0.774	0.705
LG	0.754	0.701	**0.779**	** 0.720
ADT	0.726	0.665	0.756	0.579
LMT	**0.773**	0.702	0.500	0.500

The underlined values indicate the models with AUC>0.7; the blue-coloured values denote the best model for each set of models; the double asterisked (**) values represent the best generic model

In Table 30, GRACE_CMM_Ext and CFS_CMM_Ext represent the models that have been validated externally for Malaysian and UK models, respectively.

The AUC scores obtained for the models from external validation were anticipated to be lower than those validated on internal datasets. Thus, an AUC score of 0.700 was considered acceptable when tested on externally. As presented in Table 30, the best generic models for both the Malaysian and UK datasets was CFS_CMM. CFS_CMM is a model developed based on UK dataset using LG algorithm has obtained AUC of 0.779 when validated on similar cohorts, and an AUC of 0.720 when validated on the external dataset.

8.2.3 Calibration

Based on the results derived from internal validation, the best models revealed for both the Malaysian and UK datasets were GRACE on the NB algorithm (referred to as MY_GRACE_NB) and GRACE on the LMT algorithm (referred to as UK_GRACE_LMT). In addition, the best generic model was CFS_CMM, developed on UK datasets using LG algorithm (referred to as UK_CFS_CMM _LG). As such, the BS and calibrated plots of MY_GRACE_NB, UK_GRACE_LMT and UK_CFS_CMM _LG are illustrated in Figures 15, 16, and 17, respectively. Additionally, Figure 16 portrays the BS and calibration plots for UK_CFS_CMM _LG validated on an external dataset(referred to as UK_CFS_CMM_LG_Ext).

As a whole, the BSs of the calibrated models mainly approached zero, and all of the BSs of the models were less than 0.07,indicatingthat the models were indeed well-calibrated. In fact, UK_CFS_CMM_LG obtained the best BS of 0.025, followed by UK_GRACE_LMT (BS = 0.032), UK_CFS_CMM_LG_Ext (BS = 0.062), and MY_GRACE_NB (BS = 0.063).

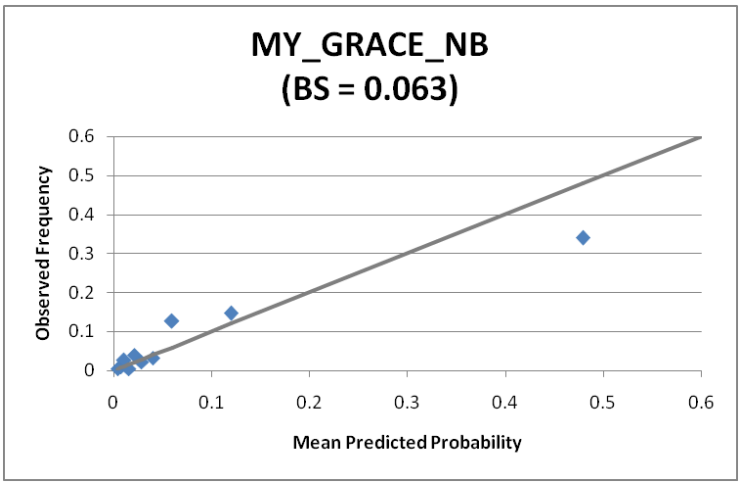


Figure 15:Calibration plot for the MY_GRACE_NB model

An obvious miscalibration was noted on bins 10 and 8 of the MY_GRACE_NB model depicted in Figure 15. This is because the model over-estimated the occurrence of the "Died" cases on bin 10, but under-estimated the same event on bin 8.

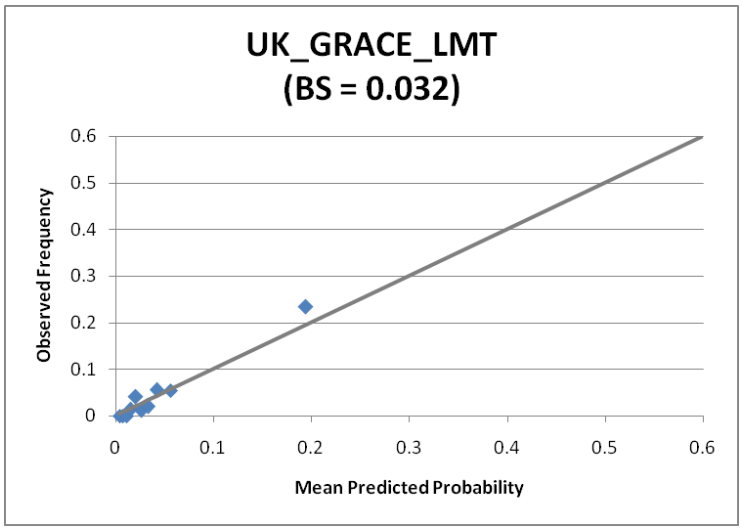


Figure 16 : Calibration plot for the UK_GRACE_LMT model

The calibration plot (Figure 16) of the UK_GRACE_LMT model depicts that most of the points for the bins were nearly on the 45-degree line, except for bin 10. Nevertheless, the distance of the point on bin 10 to the diagonal line was small. Therefore, one can claim that the UK_GRACE_LMT was calibrated significantly well.

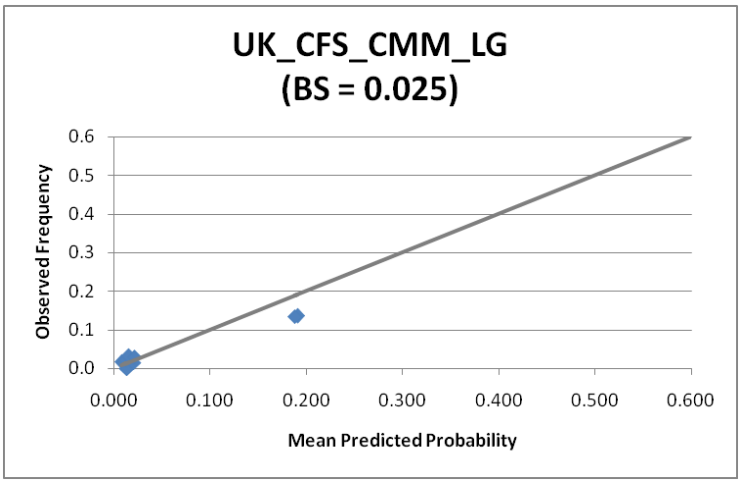


Figure 17 : Calibration plot for the UK_CFS_CMM_LG model

Likewise, for UK_GRACE_LMT, the calibration plot for the UK_CFS_CMM_LG model (Figure 17) also points out good calibration. With probability 0.02, an indicator of an over-estimated true value was present, but the difference of the estimation from the actual value was only 0.05.

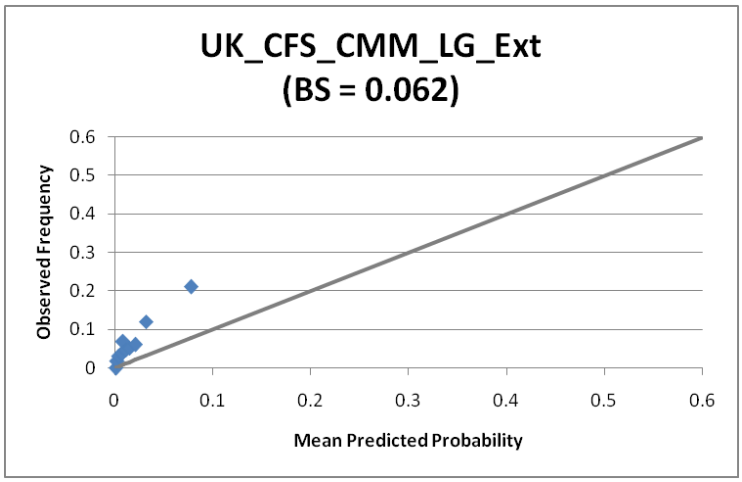


Figure 18: Calibration plot for the CFS_CMM_LG_Ext model

The calibration plot of UK_CFS_CMM_LG_Ext (Figure 18) signifies that the model under-estimated the actual output for all the bin points. The variance in mean predicted probability and actual output was 0.0009 on bin 1, 0.0149 on bin 2, 0.0251 on bin 3, 0.0314 on bin 4, 0.0595 on bin 5, 0.0362 on bin 6, 0.0373 on bin 7, 0.0395 on bin 8, 0.0892 on bin 9, and 0.1333 on bin 10. Although the model under-estimated the outcome for all bin points, the differences between the predicted and true values were not obvious.

8.3. Discussion and Conclusion

The main objective of this chapter was to identify the best model for both the Malaysian and UK cohorts. For that reason, internal and external validation was performed. The number of predictors for the best models identified from internal validation was reduced to allow the external validation processes. Other than that, calibration is another essential measure that determines the performance of a prediction model. This study measured the calibration by using BS and projected the calibration on calibration plots.

As a result, the findings concluded that the best models that predicted ACS mortality specific to the Malaysian and UK cohorts were models derived from the set of predictors of the GRACE model. Although the predictors of the model were based on those of the GRACE model, only predictors available in the Malaysian and UK datasets were incorporated into the models. As such, instead of eight predictors that present in the GRACE model (Granger et al., 2003), the best model for Malaysian dataset only incorporated five predictors, whereas the best model for UK dataset only considered four predictors. Furthermore, the Malaysian best model performed the best on NB (AUC=0.827), while the UK best model performed the best on LMT (AUC=0.847). In addition, the BSs of the models indicated that all the models were indeed well-calibrated. The calibration plots further supported the BS results as satisfactory.

On top of that, the models that displayed improvement in the model optimization process were also validated. In the previous chapter, the results showed that the application of the *random undersampling* approach on the UK dataset had improved the AUC scores of the models. Nonetheless, when the models with *random undersampling* approach were

validated, no notable enhancements were observed. The results further signified that, even with an imbalanced dataset, a good AUC (> 0.80) could still be attained.

Additionally, the external validation emphasised a common best model that predicted both the Malaysian and UK datasets. Therefore, the best generic model was constructed based on the UK cohort, using predictors extracted from *CFS* method and developed using the LG algorithm (UK_CFS_CMM_LG). Furthermore, the model was derived from a set of predictors determined by the *CFS* automated feature selection approach. In addition, the predictors of the best generic model were comprised of age, SBP, and BB taken. As a result, the model achieved an AUC=0.779 when validated on the same cohorts and an AUC=0.720 when validated on external cohorts. For external validation, the study considered an AUC of 0.70 as an acceptable and good model. Furthermore, the BS of the UK_CFS_CMM_LG model suggested good calibration measure. Nevertheless, the calibration plot of the model demonstrated under-estimated prediction for all the bin points. Nonetheless, the variance of probability predicted and true values was small.

Chapter 9: Discussion, Future Research, and Conclusion

This chapter presents the summary of the study findings and discusses the perspectives derived from the findings.

9.1. Overall Findings

9.1.1 ACS Prediction Models on ML Algorithm

This study has successfully demonstrated a practical way of constructing ACS prediction models by using ML algorithms on registry datasets. The major finding is that ML algorithms present a competitive alternative with which to build ACS prediction models. Furthermore, a number of ML algorithms have exhibited superior discriminative ability when compared to existing models developed with traditional statistical methods. For example, the models utilizing a *CSF* feature section method on the UK dataset, built using the NB, LG, MLP, FT, and LMT algorithms, achieved higher predictive power than TIMI, PURSUIT, GRACE, EMMACE, SRI, and C-ACS (Antman et al., 2000, Dorsch et al., 2001, Morrow et al., 2001, Huynh et al., 2013, Boersma et al., 2000, Granger et al., 2003).

Furthermore, the models built on an ML algorithm with predictors from an existing ACS model displayed enhanced performance, in comparison to the original model. As presented in Section 5.2.2, the best derivation models constructed for the Malaysian and UK datasets attained higher AUC values than all the 11 reviewed ACS models (Table 1). For instance, 11 out of 17 algorithms used to developed models adopting predictors from the GRACE model, based on the Malaysia dataset, had higher AUC values, i.e., c-statistics, than the GRACE model (AUC = 0.83)(Granger et al., 2003). In fact, when validated, the same model, using cohorts from the UK, attained reasonable AUC values greater than the cut-off of 0.70.

On top of that, different datasets from 2 differing regions presented varied patient characteristics due to disparities in the quality of the healthcare system, demographic diversity, lifestyle, and other factors. This variability seemed to contribute to the varying performance values among

the models. However, the datasets did have the same range of ACS, targeted the same outcomes, and had approximately similar requirements of cohorts, so the same classification algorithms resulted in rather similar performances. As noted in the findings from the STATLOG studies, the largest algorithm comparison studies on a large number of different types of datasets, there is no one best algorithm that best fits all datasets, but the same algorithms work best on datasets with similar characteristics (King et al., 1995). These findings were supported by other studies, such as Harper (2005) and Ali and Smith (2006). Thus, this study concludes that NB, BN, LG, ADT, and LMT appear to be the range of algorithms best suited for prediction modelling on ACS, and they are probably applicable to CVD, in general, and other medical datasets with similar dataset characteristics. Simple datasets characteristic measures, as outlined by Ali and Smith (2006),were made transparent in the study along the model development process as reference. The characteristics are: 1) number of predictors, 2) number of samples, 3) percentage of minority and majority classes, 4) percentage of categorical and numerical predictors, and 5) percentage of missing values. This set of algorithms (NB, BN, LG, ADT, and LMT) can be used as guideline for relatively naive medical users who wanted to attempt ML prediction modelling.

However, our findings are not consistent with a study by Potter (2007). In his study, Potter examined 56 WEKA algorithms on two breast cancer datasets. He found that no single algorithm that worked well for both datasets, even though the two datasets were similar and had a similar domain. The best classification algorithms found changed when the number of predictors differed. Even the top five algorithms were different for the two datasets. Our findings, on the other hand, even with slight differences in the actual performance values of each dataset, showed that similar domains (even with different populations and distributions of samples) and similar datasets characteristics resulted in a consistently similar set of classification algorithms. The similar performances of the algorithms were also noticed even when feature selection was applied to the datasets. In addition, this study also found that the same set of classification algorithms was not suitable for both datasets (i.e. VP, CR, Ridor, ZR, SVM, JRip, OneR, BFT,

j48, j48Graft, SC, KNN, DT, DS, RT, and REPT). Our results are consistent with Harper's (2005) study. Harper's study evaluated four classification algorithms for four different medical datasets. The findings indicated CART algorithms performed consistently well in terms of the accuracy rate, but that regression and ANN had a similar accuracy performance for almost all the datasets, and discriminant analysis (DA) as the worst algorithms for all of the datasets. In addition, King et. al (1995) found that the Bayes learner seemed to worked best on medical datasets and that NB is one of the best algorithms for our datasets. Furthermore, Wu et. al (2010)identified the LG algorithm as being better than SVM and boosting algorithms in evaluating EHR datasets. In fact, in the study, SVM was found to be the worst algorithm due to imbalanced datasets. Again, these findings are consistent with our study. In our study, LG was discovered to be one of the best algorithms, and SVM was among the non-performing algorithms, most likely due to the same reason, i.e., imbalanced datasets. Specifically related to ACS prediction modelling, our results showed that ADT was one of the best algorithms, which was also claimed by Sladojević et al.(2015)in his study.

It has been found that most highly evaluated classification algorithms in the ACS-related domain were DT, NN, SVM, and LG (Yoo et al., 2012, Liao et al., 2012, Patel and Patel, 2016). This study, on the other hand, evaluated29 ML algorithms on two datasets on the ACS domain derived from populations from different regions and with varying combinations of predictors. Furthermore, the AMIS and Serbia models, which also were compared on several WEKA algorithms, were developed from one dataset and on one identified set of predictors. Given this, it is believed that the evaluations performed in this study have indeed been thorough and extensive.

9.1.2 Data Quality

Even with adequate validation and cleaning-up processes done prior to transferring the data from hospital admission records/EHRs to the registries, data quality still appeared to be a challenge when working with this EHR-based registry data. Thus, pre-processing and data preparation was time consuming. Time was also consumed scrutinizing the attributes in the datasets so as to ascertain that only valuable attributes were selected for research analysis and model development. The study has presented the effect on model development time when dealing with issues of data quality in a dataset.

Furthermore, the study has also presented the effect on quality issues when losing a large portion of a sample. As specified in Section 4.4.1, quality issues in the datasets resulted in losing approximately 90% of the UK dataset. Since we had quite a large raw dataset, the loss did not affect the reliability of our models. But, the risk of losing a large number of observations must be considered when dealing with a dataset with quality issues that might consist mostly of EHR data or medical data, in general. This is an important note. made to encourage better quality of EHR data for further utilization of EHR data in research.

This study suggests that a good reference for data definition and description, such as data dictionary, is indeed an important asset in working with registry data. Incomplete, vague, wrong, and nil descriptions of attributes are some instances of issues discovered in the data dictionary of the studied datasets. When the registry is open for research, an extensive and detailed data dictionary should be made available, especially for users who are unfamiliar with medical data. The description of an attribute, measurement of a value, and the event or condition, complete with specific measurement metrics, have to be clearly defined. In addition, an introduction to a specific domain (e.g., ACS) and standard hospital practices would serve as added value for the researcher in better comprehending the data.

Furthermore, this study also found that the granularity of the information and how the ACS data should be stored differed between Malaysia and the UK. In fact, some information was not mutually available

in both regions. Hence, in order to attain the most advantageous prediction model, it has been essential to develop a prediction model customized by region- either by constructing a model for a specific region or updating the model in accordance to that region.

9.1.3 Predictors of ACS Models

Another major finding obtained from this study is related to the predictors for constructing ACS prediction models. In the study, in addition to producing a simpler model, ML feature selection method has demonstrated its capability in identifying a set of predictors able to construct a competitive ACS prediction model. For example, the models developed with the *CFS* feature selection method on the Malaysian dataset using the LMT algorithm had better predictive power than the TIMI, EMMACE, SRI, and C-ACS models (Antman et al., 2000, Dorsch et al., 2001, Morrow et al., 2001, Huynh et al., 2013). Even though there is no finding on a totally new predictors for ACS mortality, this different set of predictors could suggest a better ACS prediction model. Furthermore, the potential predictors resulting from ML feature selections are consistent with the existing risk factors, indicating that the ML feature selection method can identify the same risk factors as clinical trials/ medical opinions can. Thus, this study has demonstrated that the ML feature selection method could be competitive in discovering new sets of predictors for prediction modelling. In traditional clinical trial, predictors are determined by pre-selecting several potential predictors and then calculate the coefficient of the pre-selected predictors against the outcome to select significant predictors for the model. On the contrary, with the advent of the big data era, the growth of medical data is extremely rapid and sizeable, ML can be utilized for screening larger risk factor collections. These large datasets can be screened for potential predictors, as well as allowing the machine to identify the best set of predictors or even new research questions. With that, any new or vital predictors could be easily determined, and the findings can be supported and validated by scientific clinical trial.

This study concluded that the *CFS* of *filter* method is better than the two *wrapper* methods investigated in the study. However, this study does not totally agree with the finding by Hall (2000), which concluded that, in

general, the *filter* method is better than the *wrapper* method. In this study, no concrete pattern was found suggesting that the *filter* method is indeed better than the *wrapper* method. Even though the *CFS* of *filter* method resulted in better prediction models for both the Malaysian and UK datasets, the *filtersubset* of another *filter* method produced the worst models when compared to the two *wrapper* methods on the UK dataset.

In another perspective, this study showed that most of the predictors of the existing ACS prediction model are still good enough to be used as the basis of predictors in building prediction models. In Section 5.2.2, the results demonstrated that models developed adapting the set of predictors from the GRACE model were able to produce better models than the original GRACE model. In fact, for the Malaysian dataset, almost half of the algorithms out of 17 evaluated algorithms were able to construct models with better predictive performance than the original model. In addition, models developed adopting combination of predictors from nine ACS models also resulted in competitive models for both the Malaysian and UK datasets. This is an important message, suggesting that existing predictors can be adopted in developing a simple ACS model using ML customized to specific cohorts. This will save an extensive amount of model development time. This finding suggests that, despite varying characteristics of the populations and the different quality in healthcare systems between Asia and Western regions, the effect of traditional risk factors upon the outcome seemed to remain constant.

On top of that, in evaluating predictors of different clinical categories, this study has discovered that, in order to build a good ACS prediction model, the predictors must reflect a combination of information from varied phases of clinical events. As the information is varied from multiple clinical events, a better model could be developed. Nonetheless, in order to build a good model using basic or first-contact patient information, the predictors must cover at least data from the demographic, medical history, and clinical presentation categories. Moreover, this study also found predictors from the medication received before admission category (i.e. specific medicine, such as a statin or BB, that was prescribed to a patient before the ACS event) do not contribute towards improving the performance of prediction models. The

outcomes of the ML feature selection method support this finding. In all the evaluated feature selection methods, with the exception of the *wrapper* method on the Malaysian dataset, predictors from the medication received before admission category were hardly selected. Nevertheless, each dataset from a different population had its own preferred set of predictors for producing the best models. As a result, the study has found that the best sets of predictors to construct ACS models from Malaysian dataset are : 1) age, heart rate, SBP, DBP, ECG Abnormalities - T-Wave inversion, and Lvef 2) age, heart rate, SBP, killip class, ACS symptoms before admission 3) age, history of premature CVD, history of heart failure, history of lung disease, history of renal failure, heart rate, SBP, DBP, ECG Abnormalities - T-Wave inversion, ECG Abnormalities - BBB, ECG Abnormalities - Non specific, ECG Abnormalities Location - Anterior Leads : V1 and V4, ECG Abnormalities Location - Right Ventricle : ST Elevation in Lead V4R, Low-density lipoprotein cholesterol(LDL-C), FBG, Lvef, Low molecular weight heparin (LMWH) taken, Angiotensin converting enzyme (ACE)inhibitors taken, diuretics taken, and anti-arrhythmic taken . As for the UK dataset, the best sets of predictors to construct ACS models are: 1) age, BB, SBP, cardiac arrest, and reinfarction 2) age, gender, history of heart failure, on aspirin status, SBP, heart rate, cardiac arrest, ST-segment deviation of ECG 3) age, history of cerebrovascular disease, history of chronic renal failure, history of heart failure, diabetics, smoking status, aspirin status, BB, SBP, cardiac arrest, reinfarction, ECG, and tropinin assay. On the other hand, the predictors of the best generic model are age, SBP, and BB taken.

9.1.4 Misclassification Instances

In evaluating the problem that tempered the performance of the ACS model for the datasets, the study identified imbalanced datasets as the main problem. Due to this, a new approach to the *undersampling* method was introduced, i.e. *Overlapped-undersampling,* to handle the imbalanced datasets. In the *Overlapped-undersampling* approach, all the overlapped instances in the majority class were removed to achieve a fair balance distribution as existed in the minority class. This method was then compared with the existing methods for handling imbalanced datasets, such as r*andom undersampling*, *boosting*, *voting,* and using RF algorithms. The

study showed that the proposed approach made no obvious improvement over the existing approaches on the datasets, with the exception of *boosting*. Even so, the *boosting* method only worked on the UK dataset and only on the BN, ADT, and LMT algorithms. In fact, this study found that, with sufficient sample size, an imbalance dataset could be better addressed without the need for these methods. The finding is consistent with the study by Japkowicz et al.(2002). That study concluded that, with a sufficient sample size for each sub-cluster in a dataset, imbalanced datasets should not pose a problem(Japkowicz et al., 2002). This is indeed an interesting point to make for a registry dataset used for prediction modelling.

In addition, this study also discovered that overlapping instances in the minority class are yet another reason for performance degradation. Nonetheless, we believed that this reflects the underlying problem of imbalanced datasets. In an earlier study, Denil and Trappenberg (2010) had pointed to the same argument. They suggested that, when instances of imbalanced and overlapped data are present in a dataset, decline in performance could be expected. Their study evaluated the effect of overlap and imbalance issues, as well as their relationships to the size of the training set, specifically on the SVM algorithm. Our study has given a deeper perspective on overlapped instances and imbalanced datasets since our study found overlapped instances in minority class is indeed causing the problem. And, unlike Denil and Trappenberg's(2010) study, we evaluated on five ML algorithms instead of one ML algorithm.

As missing values are a major concern when developing models from a registry, a proper way of handling these missing values should be carried out. Discarding attributes with missing values is always an unwise strategy when dealing with a high-dimensionality dataset with a large number of missing data. A very limited number of complete cases or no complete cases can be achieved when trying to remove instances with missing data in this case. As described in Section 4.4.1.1, no complete instances were formed for the UK dataset, and only 318 complete instances could be extracted. Thus, as applied in this study, for a dataset that contains a large number of attributes with a large number of missing values, it is well-advised to first identify the best set of predictors for model development before

removing the instances with missing values. Also, eliminating instances with missing values does not hamper the performance of a model if the training dataset is of substantial size.

On the other hand, the study also has introduced a new approach for identifying imputation values for the missing values in a dataset, i.e., the *mean-clustering-imputation method*. Unlike the simple imputation method, which imputes the mean or the most frequent value, our imputation values were derived from clusters of the datasets. The datasets were first clustered using *Simple EM,* and the imputation value derived by calculating the mean (for numerical attributes) or the most frequent value (for categorical attributes) of each cluster. The *mean-clustering-imputation* method attained better models compared to the simple imputation method and methods embedded in specific algorithms, especially the BN, LG, and LMT algorithms. In fact, the *mean-clustering-imputation* method is more competent when more than two numerical attributes with missing values are in the dataset. However, the *mean-clustering-imputation* method is not suitable for use when the missing values are found more frequently in categorical attributes.

Finally, the study proposes a prediction model to predict misclassified instances using clinical properties as the predictors. The model was developed based on the UK dataset, using the LMT algorithm. Furthermore, the model could benefit an ACS DSS by reducing automation bias.

9.2. Main Research Contributions

The main contributions to research have been to achieve the Objectives of the study, as listed in Chapter 1:

Objective 1: This research has developed ACS mortality prediction models using DM and ML techniques that fit the Malaysian and UK datasets, and a generic dataset geared to both demographics.

Objective 2: This research has investigated ML feature selection methods and techniques for building simpler models with improved prediction power. The research also has evaluated the potency of existing sets of predictors to be adapted to other ACS registries data. Furthermore, the strength of predictors from different clinical categories in contributing to model development has also investigated.

Objective 3: This research has analysed the misclassification cases in constructing the prediction models and has identified the causes of performance degradation for the datasets. A prediction model to predict misclassified instances of the dataset using clinical information as predictors has also developed.

Objective 4: This research has investigated and evaluated ML optimization strategies to address an imbalanced dataset and missing values. The new *overlapped-undersampling* method to handle imbalanced datasets and *mean-clustering-imputation* method to handle missing values have been developed and compared with existing methods.

9.3. Limitations and Future Researches

Even though we have built competitive models specific to the Malaysian and UK datasets, and also a model that can support both cohorts, we noted that the models still need further validations on new datasets from various cohorts and settings. Particularly in the case of the Malaysian dataset, the collaboration will continue to validate the model on the latest NCVD data.

The sets of algorithms that best suit the studied datasets need to further validated on other medical datasets, as well as datasets of different domains but with similar characteristics, to further affirm generalizability of

the algorithms towards those datasets which have similar characteristics. In addition, the study has observed the effect of ML algorithms on datasets limited to only simple dataset characteristics, such as number of predictors, number of samples, percentages in the minority and majority classes, percentage of categorical and numerical predictors, and percentage of missing values in evaluating the best set of algorithms for the datasets. This work should be extended further by measuring statistical characteristics of the datasets, such as kurtosis, skewness, and correlations, as applied in the comparison of ML algorithms studies by King et al. (1995) and Ali and Smith (2006).

In the misclassification analysis study, the study only evaluated misclassified instances limited to the five best algorithms found in this study. Misclassification analysis of other popular algorithms, such as NN, RF, and DT, should provide deeper insight into the matter. In fact, the results of model development on other algorithms in the study are sufficient to be extracted and further analysed the matter.

In this study, overlapped instances in the minority class were found to be the underlying problem of the minority class. This is an important finding for future research, especially in the context of an imbalanced dataset.

Future research also should extend this work to assess the feasibility and benefits of the model in a practical clinical setting, especially in Malaysia, as the NCVD registry is now more easily accessed by the author as a result of this research . Furthermore, there is a bright opportunity to develop further a long-term ACS prediction model for the Malaysian dataset after considering the availability of PCI treatment information for ACS patients in the NCVD registry. In fact, PCI-specific prognosis models may also be constructed, thus leading to another vital contribution towards improving overall cardiac care in Malaysia.

On top of that, improvement strategies will also be brought forward to the NCVD team based on the findings and experiences gained from this study. The enhancement is targeted mainly on improving the quality of the data dictionary and supporting documents to better prepare the registry for adverse body of researchers.

In addition, it is suggested that the NCVD dataset to be made publicly available. For instance, with strict de-identification of patients, the sample of the dataset can be shared with theUCIMLrepository(https://archive.ics.uci.edu/ml/datasets.html). This allows the dataset to be accessed by not only to medical researchers, but also by other groups, such as DM and ML communities or those involved in big data studies. Moreover, this offers an opportunity for new and unexpected research questions, apart from stimulating innovative ideas. For instance, ML researchers can look further into overlapped instances in minority classes so as to address the underlying problem of an imbalanced dataset.

9.4. Conclusion

The value of developing ACS prediction models using ML has been successfully presented in this study. Competitive ACS prediction models using ML have been developed by demonstrating the practical application of different ML algorithms and methods. Evaluation of predictors of existing ACS models, and of different clinical categories, has provided insight into how to construct better ACS prediction models. Misclassification analysis has identified the underlying problem of imbalanced datasets as overlapped instances in the minority classes. In addition, missing values were also found to be one of the critical problems in misclassified instances in the datasets. The proposed correction method, i.e., the *overlapped-undersampling* method, used to handle imbalanced datasets failed to improve model performance. Other existing methods of handling imbalanced datasets, such as bagging, *random undersampling*, and *voting*, also seemed to fail in improving the overall performance of the models. Nonetheless, having a larger sample size was found to be a convincingly better way to tackle issue of imbalanced datasets. Furthermore, the proposed *mean-clustering-imputation* method for filling in missing values displayed improvement in terms of model performance in comparison to the simple imputation method and the algorithms' built-in methods. However, removing instances with missing values after feature selection is indeed the best way of handling missing values for the datasets.

List of References

A.SUDHA, P.GAYATHRI & N.JAISANKAR 2012. Effective Analysis and Predictive Model of Stroke Disease using Classification Methods. *International Journal of Computer Applications (0975 – 8887),* Volume 43– No.14.

ABEGUNDE, D. O., MATHERS, C. D., ADAM, T., ORTEGON, M. & STRONG, K. 2007. The burden and costs of chronic diseases in low-income and middle-income countries. *The Lancet,* 370**,** 1929-1938.

AHMAD, W. A. W., ZAMBAHARI, R., ISMAIL, O., SINNADURAI, J., ROSMAN, A., PIAW, C. S., ABIDIN, I. Z. & KUI-HIAN, S. 2011. Malaysian National Cardiovascular Disease Database (NCVD)–Acute Coronary Syndrome (ACS) registry: How are we different? *CVD Prevention and Control,* 6**,** 81-89.

ALI, S. & SMITH, K. A. 2006. On learning algorithm selection for classification. *Applied Soft Computing,* 6**,** 119-138.

ALLENDER, S., SCARBOROUGH, P., PETO, V., RAYNER, M., LEAL, J., LUENGO-FERNANDEZ, R. & GRAY, A. 2008. European cardiovascular disease statistics.

AMBLER, G., OMAR, R. Z. & ROYSTON, P. 2007. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical methods in medical research,* 16**,** 277-298.

ANDERSON, J. L., ADAMS, C. D., ANTMAN, E. M., BRIDGES, C. R., CALIFF, R. M., CASEY, D. E., CHAVEY, W. E., FESMIRE, F. M., HOCHMAN, J. S. & LEVIN, T. N. 2007. ACC/AHA 2007 guidelines for the management of patients with unstable angina/non–ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines for the Management of Patients With Unstable Angina/Non–ST-Elevation Myocardial Infarction) developed in collaboration with the American College of Emergency Physicians, the Society for Cardiovascular Angiography and Interventions, and the Society of Thoracic Surgeons endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation and the Society for Academic Emergency Medicine. *Journal of the American College of Cardiology,* 50**,** e1-e157.

ANTMAN, E. M., COHEN, M., BERNINK, P. J. L. M., MCCABE, C. H., HORACEK, T., PAPUCHIS, G., MAUTNER, B., CORBALAN, R., RADLEY, D. & BRAUNWALD, E. 2000. The TIMI risk score for unstable angina/non–ST elevation MI. *JAMA: the journal of the American Medical Association,* 284**,** 835-842.

ANTMAN, E. M., MCCABE, C. H., GURFINKEL, E. P., TURPIE, A. G., BERNINK, P. J., SALEIN, D., DE LUNA, A. B., FOX, K.,

LABLANCHE, J.-M. & RADLEY, D. 1999. Enoxaparin Prevents Death and Cardiac Ischemic Events in Unstable Angina/Non–Q-Wave Myocardial Infarction Results of the Thrombolysis In Myocardial Infarction (TIMI) 11B Trial. *Circulation,* 100**,** 1593-1601.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F. & NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16**,** 412-424.

BARAKAT, N. H., BRADLEY, A. P. & BARAKAT, M. N. H. 2010. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *Ieee Transactions on Information Technology in Biomedicine,* 14**,** 1114-1120.

BASSAND, J.-P., HAMM, C. W., ARDISSINO, D., BOERSMA, E., BUDAJ, A., FERNANDEZ-AVILES, F., FOX, K., HASDAI, D., OHMAN, E. M. & WALLENTIN, L. 2008. Guidelines for the diagnosis and treatment of non-ST-segment elevation acute coronary syndromes. *Revista portuguesa de cardiologia: orgão oficial da Sociedade Portuguesa de Cardiologia= Portuguese journal of cardiology: an official journal of the Portuguese Society of Cardiology,* 27**,** 1063.

BASSAND, J.-P., HAMM, C. W., ARDISSINO, D., BOERSMA, E., BUDAJ, A., FERNÁNDEZ-AVILÉS, F., FOX, K. A., HASDAI, D., OHMAN, E. M. & WALLENTIN, L. 2007. Guidelines for the diagnosis and treatment of non-ST-segment elevation acute coronary syndromes The Task Force for the Diagnosis and Treatment of Non-ST-Segment Elevation Acute Coronary Syndromes of the European Society of Cardiology. *European Heart Journal,* 28**,** 1598-1660.

BATE, A., LINDQUIST, M. & EDWARDS, I. 2008. The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. *Fundamental & Clinical Pharmacology,* 22**,** 127-140.

BATISTA, G. E., PRATI, R. C. & MONARD, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter,* 6**,** 20-29.

BAUER, E. & KOHAVI, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning,* 36**,** 105-139.

BEALE, T. 2005. The Health Record-Why is it so hard? *IMIA Yearbook of Medical Informatics,* 2005**,** 301-304.

BELEITES, C., NEUGEBAUER, U., BOCKLITZ, T., KRAFFT, C. & POPP, J. 2013. Sample size planning for classification models. *Analytica chimica acta,* 760**,** 25-33.

BERMEJO, P., DE LA OSSA, L., GÁMEZ, J. A. & PUERTA, J. M. 2012. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems,* 25**,** 35-44.

BLUM, A. L. & LANGLEY, P. 1997. Selection of relevant features and examples in machine learning. *Artificial intelligence,* 97**,** 245-271.

BODY, R., CARLEY, S., MCDOWELL, G., FERGUSON, J. & MACKWAY-JONES, K. 2009. Can a modified thrombolysis in myocardial infarction risk score outperform the original for risk stratifying emergency department patients with chest pain? *Emergency Medicine Journal,* 26**,** 95-99.

BOERSMA, E., PIEPER, K. S., STEYERBERG, E. W., WILCOX, R. G., CHANG, W.-C., LEE, K. L., AKKERHUIS, K. M., HARRINGTON, R. A., DECKERS, J. W. & ARMSTRONG, P. W. 2000. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation results from an international trial of 9461 patients. *Circulation,* 101**,** 2557-2567.

BOUCKAERT, R. R. 2008. *Bayesian Network Classifiers in Weka* [Online]. The University of Waikato - Department of Computer Science. Available: http://www.cs.waikato.ac.nz/~remco/weka.bn.pdf [Accessed 12 September 20`7 2017].

BOUWMEESTER, W., ZUITHOFF, N. P., MALLETT, S., GEERLINGS, M. I., VERGOUWE, Y., STEYERBERG, E. W., ALTMAN, D. G. & MOONS, K. G. 2012. Reporting and methods in clinical prediction research: a systematic review. *PLoS medicine,* 9**,** e1001221.

BRADLEY, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition,* 30**,** 1145-1159.

BREIMAN, L. 1996. Bagging predictors. *Machine learning,* 24**,** 123-140.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

BRODLEY, C. E. & FRIEDL, M. A. Identifying and eliminating mislabeled training instances. Proceedings of the National Conference on Artificial Intelligence, 1996. 799-805.

CHASE, M., ROBEY, J. L., ZOGBY, K. E., SEASE, K. L., SHOFER, F. S. & HOLLANDER, J. E. 2006. Prospective validation of the Thrombolysis in Myocardial Infarction Risk Score in the emergency department chest pain population. *Annals of emergency medicine,* 48**,** 252-259.

CHAWLA, N. V. 2010. Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook***,** 875-886.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. & KEGELMEYER, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* 16**,** 321-357.

CHAZARD, E., FICHEUR, G., BERNONVILLE, S., LUYCKX, M. & BEUSCART, R. 2011. Data Mining to Generate Adverse Drug Events Detection Rules. *Ieee Transactions on Information Technology in Biomedicine,* 15**,** 823-830.

CHIN, S., JEYAINDRAN, S., AZHARI, R., WAN AZMAN, W., OMAR, I., ROBAAYAH, Z. & SIM, K. 2008. Acute coronary syndrome (ACS) registry--leading the charge for National Cardiovascular Disease (NCVD) Database. *Med J Malaysia,* 63**,** 29-36.

CHOUDHARY, A., HARDING, J. & TIWARI, M. 2009. Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing,* 20**,** 501-521.

CIOS, K. J. & MOORE, G. W. 2002. Uniqueness of medical data mining. *Artificial Intelligence in Medicine,* 26**,** 1-24.

CLAHRC FOR LEEDS, Y., AND BRADFORD. *Vascular Disease: Improve-PC* [Online]. NIHR. Available: http://www.clahrc-lyb.nihr.ac.uk/research-and-development/improve-pc/ [Accessed 18/11/2014 2014].

COOK, N. R. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation,* 115**,** 928-935.

COONEY, M. T., DUDINA, A. L. & GRAHAM, I. M. 2009. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. *Journal of the American College of Cardiology,* 54**,** 1209-1227.

D'ASCENZO, F., BIONDI-ZOCCAI, G., MORETTI, C., BOLLATI, M., OMEDÈ, P., SCIUTO, F., PRESUTTI, D. G., MODENA, M. G., GASPARINI, M. & REED, M. J. 2012. TIMI, GRACE and alternative risk scores in Acute Coronary Syndromes: a meta-analysis of 40 derivation studies on 216,552 patients and of 42 validation studies on 31,625 patients. *Contemporary clinical trials,* 33**,** 507-514.

DANGARE, C. S. & APTE, S. S. 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. *International Journal of Computer Applications,* 47**,** 44-48.

DAS, S. Filters, wrappers and a boosting-based hybrid for feature selection. ICML, 2001. Citeseer, 74-81.

DE ARAÚJO GONÇALVES, P., FERREIRA, J., AGUIAR, C. & SEABRA-GOMES, R. 2005. TIMI, PURSUIT, and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTE-ACS. *European heart journal,* 26**,** 865-872.

DEL FIOL, G. & HAUG, P. J. 2009. Classification models for the prediction of clinicians' information needs. *Journal of Biomedical Informatics,* 42**,** 82-89.

DELEN, D. 2009. Analysis of cancer data: a data mining approach. *Expert Systems,* 26**,** 100-112.

DELEN, D., COGDELL, D. & KASAP, N. 2012a. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting,* 28**,** 543-552.

DELEN, D., OZTEKIN, A. & KONG, Z. 2010. A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artificial Intelligence in Medicine,* 49**,** 33-42.

DELEN, D., OZTEKIN, A. & TOMAK, L. 2012b. An analytic approach to better understanding and management of coronary surgeries. *Decision Support Systems,* 52**,** 698-705.

DELEN, D., WALKER, G. & KADAM, A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine,* 34**,** 113-127.

DENIL, M. & TRAPPENBERG, T. Overlap versus imbalance. Canadian Conference on Artificial Intelligence, 2010. Springer, 220-231.

DORSCH, M., LAWRANCE, R., SAPSFORD, R., OLDHAM, J., GREENWOOD, D., JACKSON, B., MORRELL, C., BALL, S., ROBINSON, M. & HALL, A. 2001. A simple benchmark for evaluating quality of care of patients following acute myocardial infarction. *Heart,* 86**,** 150-154.

FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern recognition letters,* 27**,** 861-874.

FAYYAD, U., PIATETSKYSHAPIRO, G. & SMYTH, P. 1996. From data mining to knowledge discovery in databases. *Ai Magazine,* 17**,** 37-54.

FONAROW, G. C., ADAMS, K. F., ABRAHAM, W. T., YANCY, C. W., BOSCARDIN, W. J. & COMMITTEE, A. S. A. 2005. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *Jama,* 293**,** 572-580.

FREUND, Y. & MASON, L. The alternating decision tree learning algorithm. icml, 1999. 124-133.

FREUND, Y. & SCHAPIRE, R. E. Experiments with a new boosting algorithm. icml, 1996. 148-156.

GALAR, M., FERNANDEZ, A., BARRENECHEA, E., BUSTINCE, H. & HERRERA, F. 2012. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42**,** 463-484.

GALE, C. P., MANDA, S., BATIN, P. D., WESTON, C. F., BIRKHEAD, J. S. & HALL, A. S. 2008a. Predictors of in-hospital mortality for patients admitted with ST-elevation myocardial infarction: a real-world study using the Myocardial Infarction National Audit Project (MINAP) database. *Heart,* 94**,** 1407-1412.

GALE, C. P., MANDA, S. O., WESTON, C. F., BIRKHEAD, J. S., BATIN, P. D. & HALL, A. S. 2008b. Evaluation of risk scores for risk stratification of acute coronary syndromes in the Myocardial Infarction National Audit Project (MINAP) database. *Heart*.

GAZIANO, T. A., BITTON, A., ANAND, S., ABRAHAMS-GESSEL, S. & MURPHY, A. 2010. Growing epidemic of coronary heart disease in low-and middle-income countries. *Current problems in cardiology,* 35**,** 72-115.

GIBERT, K., SANCHEZ-MARRE, M. & CODINA, V. 2010. *Choosing the right data mining technique: classification of methods and intelligent recommendation.* International Environmental Modelling and Software Society.

GIUGLIANO, R., LLEVADOT, J., WILCOX, R., GURFINKEL, E., MCCABE, C., CHARLESWORTH, A., THOMPSON, S., ANTMAN, E. & BRAUNWALD FOR THE IN TIME II INVESTIGATORS, E. 2001. Geographic variation in patient and hospital characteristics, management, and clinical outcomes in ST-elevation myocardial infarction treated with fibrinolysis. Results from InTIME-II. *European Heart Journal,* 22**,** 1702-1715.

GLIKLICH, R. E., DREYER, N. A. & LEAVY, M. B. 2014. Interfacing registries with electronic health records.

GRANGER, C. B., GOLDBERG, R. J., DABBOUS, O., PIEPER, K. S., EAGLE, K. A., CANNON, C. P., VAN DE WERF, F., AVEZUM, A., GOODMAN, S. G. & FLATHER, M. D. 2003. Predictors of hospital mortality in the global registry of acute coronary events. *Archives of Internal Medicine,* 163**,** 2345.

GREEN, M., BJOERK, J., FORBERG, J., EKELUND, U., EDENBRANDT, L. & OHLSSON, M. 2006a. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine,* 38**,** 305-318.

GREEN, M., BJÖRK, J., FORBERG, J., EKELUND, U., EDENBRANDT, L. & OHLSSON, M. 2006b. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial intelligence in medicine,* 38**,** 305-318.

GRZYMALA-BUSSE, J. & HU, M. A comparison of several approaches to missing attribute values in data mining. Rough sets and current trends in computing, 2001. Springer, 378-385.

GUYON, I. & ELISSEEFF, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research,* 3**,** 1157-1182.

GUYON, I., WESTON, J., BARNHILL, S. & VAPNIK, V. 2002. Gene selection for cancer classification using support vector machines. *Machine learning,* 46**,** 389-422.

HALL, M. 2007. A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems,* 20**,** 120-126.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter,* 11**,** 10-18.

HALL, M. A. 2000. Correlation-based feature selection of discrete and numeric class machine learning.

HALL, M. A. & SMITH, L. A. 1998. Practical feature subset selection for machine learning.

HAN, H., WANG, W.-Y. & MAO, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing***,** 878-887.

HAN, J. & KAMBER, M. 2001. *Data mining : concepts and techniques,* San Francisco Morgan Kaufmann Publishers.

HARPER, P. R. 2005. A review and comparison of classification algorithms for medical decision making. *Health Policy,* 71**,** 315-331.

HAYRINEN, K., SARANTO, K. & NYKANEN, P. 2008. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics,* 77**,** 291-304.

HESS, E. P., AGARWAL, D., CHANDRA, S., MURAD, M. H., ERWIN, P. J., HOLLANDER, J. E., MONTORI, V. M. & STIELL, I. G. 2010. Diagnostic accuracy of the TIMI risk score in patients with chest pain in the emergency department: a meta-analysis. *Canadian Medical Association Journal,* 182**,** 1039-1044.

HOUSE, P. A., HILL, D. K., MURRAY, D. J., HONEY, D. S., CRAIGS, M. C., WARD, D. V., GALE, D. C. & BROCH, D. J. 2011. Improving Prevention of Vascular Events in Primary Care : IMPROVE-PC CLAHRC Vascular Theme

Cardiovascular Healthcare Information Linkage Study Protocol.: National Institue for Health Research (NIHR).

HRUSCHKA, E. R., HRUSCHKA, E. R. & EBECKEN, N. F. Towards efficient imputation by nearest-neighbors: a clustering-based approach. Australasian Joint Conference on Artificial Intelligence, 2004. Springer, 513-525.

HU, D., HUANG, Z., CHAN, T.-M., DONG, W., LU, X. & DUAN, H. 2016. Utilizing Chinese Admission Records for MACE Prediction of Acute Coronary Syndrome. *International Journal of Environmental Research and Public Health,* 13**,** 912.

HUANG, Y., MCCULLAGH, P., BLACK, N. & HARPER, R. 2004. Evaluation of Outcome Prediction for a Clinical Diabetes Database. *Knowledge Exploration in Life Science Informatics***,** 181-190.

HUYNH, T., KOUZ, S., YAN, A., DANCHIN, N., LOUGHLIN, J. O., SCHAMPAERT, E., YAN, R., RINFRET, S., TARDIF, J.-C. & EISENBERG, M. J. 2013. Canada Acute Coronary Syndrome Risk Score: A new risk score for early prognostication in acute coronary syndromes. *American heart journal,* 166**,** 58-63.

IAVINDRASANA, J., COHEN, G., DEPEURSINGE, A., MÜLLER, H., MEYER, R. & GEISSBUHLER, A. 2009. Clinical data mining: a review. *Yearb Med Inform,* 2009**,** 121-133.

INVESTIGATORS, G. 1993. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl j Med,* 1993**,** 673-682.

ISILAK, Z., KARDESOGLU, E., APARCI, M., UZ, O., YALCIN, M., YIGINER, O., CINGOZBAY, B. Y. & UZUN, M. 2012. Comparison of clinical risk assessment systems in predicting three− vessel coronary artery disease and angiographic culprit lesion in patients with non− ST segment elevated myocardial infarction/unstable angina pectoris. *Kardiologia Polska (Polish Heart Journal),* 70**,** 242-250.

JAAFAR, J., ATWELL, E., JOHNSON, O., CLAMP, S. & AHMAD, W. A. W. 2013. Evaluation of Machine Learning Techniques in Predicting Acute Coronary Syndrome Outcome. *Research and Development in Intelligent Systems XXX.* Springer.

JAPKOWICZ, NATHALIE STEPHEN & SHAJU 2002. The class imbalance problem: A systematic study. *Intelligent data analysis,* 6**,** 429-449.

JAPKOWICZ, N. Learning from imbalanced data sets: a comparison of various strategies. AAAI workshop on learning from imbalanced data sets, 2000.

JO, T. & JAPKOWICZ, N. 2004. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter,* 6**,** 40-49.

JOHN, G. H., KOHAVI, R. & PFLEGER, K. Irrelevant features and the subset selection problem. Machine learning: proceedings of the eleventh international conference, 1994. 121-129.

JOHN, G. H. & LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995. Morgan Kaufmann Publishers Inc., 338-345.

JOVIĆ, A., BRKIĆ, K. & BOGUNOVIĆ, N. A review of feature selection methods with applications. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, 2015. IEEE, 1200-1205.

KANTARDZIC, M. M. & ZURADA, J. 2005. *Next generation of data-mining applications,* Hoboken, N.J., Wiley-Interscience.

KARAOLIS, M. A., MOUTIRIS, J. A., HADJIPANAYI, D. & PATTICHIS, C. S. 2010. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions on information technology in biomedicine,* 14**,** 559-566.

KAREGOWDA, A. G., MANJUNATH, A. & JAYARAM, M. 2010. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management,* 2**,** 271-277.

KHALILIA, M., CHAKRABORTY, S. & POPESCU, M. 2011. Predicting disease risks from highly imbalanced data using random forest. *Bmc Medical Informatics and Decision Making,* 11**,** 51.

KHAN, S. Q., NARAYAN, H., NG, K. H., DHILLON, O. S., KELLY, D., QUINN, P., SQUIRE, I. B., DAVIES, J. E. & NG, L. L. 2009. N-terminal pro-B-type natriuretic peptide complements the GRACE risk score in predicting early and late mortality following acute coronary syndrome. *Clinical Science,* 117**,** 31-39.

KHOSHGOFTAAR, T. M., SELIYA, N. & GAO, K. Rule-based noise detection for software measurement data. Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on, 2004. IEEE, 302-307.

KHOSLA, A., CAO, Y., LIN, C. C. Y., CHIU, H. K., HU, J. & LEE, H. An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. ACM, 183-192.

KING, R. D., FENG, C. & SUTHERLAND, A. 1995. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal,* 9**,** 289-333.

KITTLER, J., HATEF, M., DUIN, R. P. & MATAS, J. 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence,* 20**,** 226-239.

KOH, H. C. & TAN, G. 2011. Data mining applications in healthcare. *Journal of Healthcare Information Management—Vol,* 19**,** 65.

KOHAVI, R. & JOHN, G. H. 1997. Wrappers for feature subset selection. *Artificial intelligence,* 97**,** 273-324.

KOTSIANTIS, S., KANELLOPOULOS, D. & PINTELAS, P. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering,* 30**,** 25-36.

KOTSIANTIS, S. B., ZAHARAKIS, I. & PINTELAS, P. 2007. Supervised machine learning: A review of classification techniques.

KUBAT, M. & MATWIN, S. Addressing the curse of imbalanced training sets: one-sided selection. ICML, 1997. Nashville, USA, 179-186.

KUMAR, R. & INDRAYAN, A. 2011. Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics,* 48**,** 277-287.

KURZ, D. J., BERNSTEIN, A., HUNT, K., RADOVANOVIC, D., ERNE, P., SIUDAK, Z. & BERTEL, O. 2009. Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *Heart,* 95**,** 662-668.

LANDWEHR, N., HALL, M. & FRANK, E. 2005. Logistic model trees. *Machine learning,* 59**,** 161-205.

LAVESSON, N., HALLING, A., FREITAG, M., ODEBERG, J., ODEBERG, H. & DAVIDSSON, P. Classifying the severity of an acute coronary syndrome by mining patient data. The Swedish AI Society Workshop May 27-28; 2009 IDA; Linköping University, 2009. Linköping University Electronic Press, 55-63.

LE CESSIE, S. & VAN HOUWELINGEN, J. C. 1992. Ridge estimators in logistic regression. *Applied statistics***,** 191-201.

LEE, K. L., WOODLIEF, L. H., TOPOL, E. J., WEAVER, W. D., BETRIU, A., COL, J., SIMOONS, M., AYLWARD, P., VAN DE WERF, F. & CALIFF, R. M. 1995. Predictors of 30-Day Mortality in the Era of Reperfusion for Acute Myocardial Infarction Results From an International Trial of 41 021 Patients. *Circulation,* 91**,** 1659-1668.

LEE, Y.-H., BANG, H. & KIM, D. J. 2016. How to establish clinical prediction models. *Endocrinology and Metabolism,* 31**,** 38-44.

LIAO, S.-H., CHU, P.-H. & HSIAO, P.-Y. 2012. Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications,* 39**,** 11303-11311.

LIU, X.-Y., WU, J. & ZHOU, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* 39**,** 539-550.

LLOYD-JONES, D., ADAMS, R., CARNETHON, M., DE SIMONE, G., FERGUSON, T. B., FLEGAL, K., FORD, E., FURIE, K., GO, A. & GREENLUND, K. 2009. Heart disease and stroke statistics—2009 update. *Circulation,* 119**,** e21-e181.

LLOYD-JONES, D. M. 2010. Cardiovascular Risk Prediction Basic Concepts, Current Status, and Future Directions. *Circulation,* 121**,** 1768-1777.

LOBO, J. M., JIMÉNEZ-VALVERDE, A. & REAL, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography,* 17**,** 145-151.

LÓPEZ, V., FERNÁNDEZ, A., GARCÍA, S., PALADE, V. & HERRERA, F. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences,* 250**,** 113-141.

MACKAY, J., MENSAH, G. A., MENDIS, S. & GREENLUND, K. 2004. *The atlas of heart disease and stroke*, World Health Organization.

MALDONADO, S. & WEBER, R. 2009. A wrapper method for feature selection using support vector machines. *Information Sciences,* 179**,** 2208-2217.

MASIC, I., MIOKOVIC, M. & MUHAMEDAGIC, B. 2008. Evidence based medicine–new approaches and challenges. *Acta Informatica Medica,* 16**,** 219.

MORRIS, A. C., CAESAR, D., GRAY, S. & GRAY, A. 2006. TIMI risk score accurately risk stratifies patients with undifferentiated chest pain presenting to an emergency department. *Heart,* 92**,** 1333-1334.

MORROW, D. A., ANTMAN, E. M., CHARLESWORTH, A., CAIRNS, R., MURPHY, S. A., DE LEMOS, J. A., GIUGLIANO, R. P., MCCABE, C. H. & BRAUNWALD, E. 2000. TIMI risk score for ST-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation. *Circulation,* 102**,** 2031-2037.

MORROW, D. A., ANTMAN, E. M., GIUGLIANO, R. P., CAIRNS, R., CHARLESWORTH, A., MURPHY, S. A., DE LEMOS, J. A., MCCABE, C. H. & BRAUNWALD, E. 2001. A simple risk index for rapid initial triage of patients with ST-elevation myocardial infarction: an InTIME II substudy. *The Lancet,* 358**,** 1571-1575.

MUKHERJEE, S., TAMAYO, P., ROGERS, S., RIFKIN, R., ENGLE, A., CAMPBELL, C., GOLUB, T. R. & MESIROV, J. P. 2003. Estimating dataset size requirements for classifying DNA microarray data. *Journal of computational biology,* 10**,** 119-142.

NISBET, R., MINER, G. & ELDER IV, J. 2009. *Handbook of statistical analysis and data mining applications*, Academic Press.

OVERBAUGH, K. J. 2009. Acute coronary syndrome. *AJN The American Journal of Nursing,* 109**,** 42-52.

PATEL, S. & PATEL, H. 2016. Survey of data mining techniques used in healthcare domain. *International Journal of Information,* 6.

PEDERSEN, A. B., MIKKELSEN, E. M., CRONIN-FENTON, D., KRISTENSEN, N. R., PHAM, T. M., PEDERSEN, L. & PETERSEN, I. 2017. Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology,* 9**,** 157.

PEPE, M. S. 2003. *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, USA.

PFISTERER, M., COX, J. L., GRANGER, C. B., BRENER, S. J., NAYLOR, C. D., CALIFF, R. M., VAN DE WERF, F., STEBBINS, A. L., LEE, K. L. & TOPOL, E. J. 1998. Atenolol use and clinical outcomes after thrombolysis for acute myocardial infarction: the GUSTO-I experience. *Journal of the American College of Cardiology,* 32**,** 634-640.

PHILIP I. AARONSON & WARD., J. P. T. 2007. *The cardiovascular system at a glance*, Malden, Mass. : Blackwell, 2007.

POGORELC, B., BOSNIĆ, Z. & GAMS, M. 2012. Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimedia Tools and Applications***,** 1-22.

POTTER, R. Comparison of Classification Algorithms Applied to Breast Cancer Diagnosis and Prognosis. Industrial Conference on Data Mining-Posters and Workshops, 2007. 40-49.

RAHMAN, M. M. & DAVIS, D. Cluster based under-sampling for unbalanced cardiovascular data. Proceedings of the World Congress on Engineering, 2013. 3-5.

RAMENTOL, E., CABALLERO, Y., BELLO, R. & HERRERA, F. 2012. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and information systems,* 33**,** 245-265.

ROTH, G. A., FOROUZANFAR, M. H., MORAN, A. E., BARBER, R., NGUYEN, G., FEIGIN, V. L., NAGHAVI, M., MENSAH, G. A. & MURRAY, C. J. 2015. Demographic and epidemiologic drivers of global cardiovascular mortality. *New England Journal of Medicine,* 372**,** 1333-1341.

SAEYS, Y., INZA, I. & LARRAÑAGA, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics,* 23**,** 2507-2517.

SAMI, A. 2006. Obstacles and misunderstandings facing medical data mining. *In:* LI, X. Z. O. R. L. Z. H. (ed.) *Advanced Data Mining and Applications, Proceedings.*

SAMPSON, D. L., PARKER, T. J., UPTON, Z. & HURST, C. P. 2011. A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches. *Plos One,* 6.

SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J. & FOLLECO, A. 2014. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences,* 259**,** 571-595.

SHILLABEER, A. & RODDICK, J. F. Establishing a lineage for medical knowledge discovery. 2007. Australian Computer Society, Inc., 29-37.

SIGN 2007. Risk estimation and the prevention of cardiovascular disease : A national clinical guideline. *In:* SCOTLAND, N. (ed.). SIGN (Scottish Intercollegiate Guidelines Network).

SIGN 2007 (Updated 2013). Acute coronary syndromes : A national clinical guideline. Edinburgh, UK: Scottish Intercollegiate Guidelines Network.

SITAR-TAUT, A., ZDRENGHEA, D., POP, D. & SITAR-TAUT, D. 2009. Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation. *Journal of Applied Computer Science & Mathematics*.

SLADOJEVIĆ, M., ČANKOVIĆ, M., ČEMERLIĆ, S., MIHAJLOVIĆ, B., AĐIĆ, F. & JARAKOVIĆ, M. 2015. Data mining approach for in-hospital treatment outcome in patients with acute coronary syndrome. *Medicinski pregled,* 68**,** 157-161.

SMITH, J. N., NEGRELLI, J. M., MANEK, M. B., HAWES, E. M. & VIERA, A. J. 2015. Diagnosis and management of acute coronary syndrome: an evidence-based update. *The Journal of the American Board of Family Medicine,* 28**,** 283-293.

SMITH, M. R. 2009. An empirical study of instance hardness.

SMITH, M. R. & MARTINEZ, T. Improving classification accuracy by identifying and removing instances that should be misclassified. Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011. IEEE, 2690-2697.

STEFANOWSKI, J. 2013. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. *Emerging paradigms in machine learning.* Springer.

STEYERBERG, E. 2008. *Clinical prediction models: a practical approach to development, validation, and updating*, Springer Science & Business Media.

STEYERBERG, E. W. 2009. *Clinical prediction models: a practical approach to development, validation, and updating*, Springer.

STOLBA, N. & TJOA, A. M. 2006. The Relevance of Data Warehousing and Data Mining in the Field of Evidence-based Medicine to Support Healthcare Decision Making. *In:* ARDIL, C. (ed.) *Proceedings of World Academy of Science, Engineering and Technology, Vol 11.*

SU, X., KHOSHGOFTAAR, T. M. & GREINER, R. Using imputation techniques to help learn accurate classifiers. Tools with Artificial

Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on, 2008. IEEE, 437-444.

SWALES, J. D. J. D. & P. DE BONO, D. 1993. *Cardiovascular Risk Factors* London ; New York : . Gower Medical Pub.

TAN, F. 2007. Improving feature selection techniques for machine learning.

TENORIO, J. M., HUMMEL, A. D., COHRS, F. M., SDEPANIAN, V. L., PISA, I. T. & MARINE, H. D. F. 2011. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. *International Journal of Medical Informatics,* 80**,** 793-802.

THAM, C., HENG, C. & CHIN, W. 2003. Predicting risk of coronary artery disease from DNA microarray-based genotyping using neural networks and other statistical analysis tool. *Journal of bioinformatics and computational biology,* 1**,** 521-539.

THONGKAM, J., XU, G., ZHANG, Y. & HUANG, F. 2008. Support vector machine for outlier detection in breast cancer survivability prediction. *Advanced Web and NetworkTechnologies, and Applications***,** 99-109.

TOMAR, D. & AGARWAL, S. 2013. A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology,* 5**,** 241-266.

VAN CALSTER, B., NIEBOER, D., VERGOUWE, Y., PENCINA, M. J. & STEYERBERG, E. W. 2015. A calibration hierarchy for risk models: strong calibration occurs only in utopia.

VAN HULSE, J., KHOSHGOFTAAR, T. M. & NAPOLITANO, A. Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning, 2007. ACM, 935-942.

VANHOUTEN, J. P., STARMER, J. M., LORENZI, N. M., MARON, D. J. & LASKO, T. A. Machine Learning for Risk Prediction of Acute Coronary Syndrome. AMIA Annual Symposium Proceedings, 2014. American Medical Informatics Association, 1940.

VAPNIK, V. 1998. *Statistical learning theory. 1998*, Wiley, New York.

VINTERBO, S. & OHNO-MACHADO, L. A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. Proceedings of the AMIA Symposium, 1999. American Medical Informatics Association, 984.

VISA, S. & RALESCU, A. Issues in mining imbalanced data sets-a review paper. Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, 2005. sn, 67-73.

WANG, Z., SHAH, A. D., TATE, A. R., DENAXAS, S., SHAWE-TAYLOR, J. & HEMINGWAY, H. 2012. Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. *Plos One,* 7.

WEISS, G. M. The impact of small disjuncts on classifier learning. Data Mining, 2010. Springer, 193-226.

WEISS, S. M. & INDURKHYA., N. 1998. *Predictive data mining : a practical guide,* San Francisco, California, Morgan Kaufmann Publishers.

WESTON, J., MUKHERJEE, S., CHAPELLE, O., PONTIL, M., POGGIO, T. & VAPNIK, V. Feature selection for SVMs. Advances in neural information processing systems, 2001. 668-674.

WITTEN, I. H., FRANK., E. & HALL, M. A. 2005. *Data mining : practical machine learning tools and techniques,* Amsterdam ; London :, Elsevier, c2005.

WORKMAN & A, T. 2013. *Engaging patients in information sharing and data collection: the role of patient-powered registries and research networks*, Agency for Healthcare Research and Quality (US), Rockville (MD).

WU, J. L., ROY, J. & STEWART, W. F. 2010. Prediction Modeling Using EHR Data Challenges, Strategies, and a Comparison of Machine Learning Approaches. *Medical Care,* 48**,** S106-S113.

YANG, C.-H., CHUANG, L.-Y. & YANG, C. H. 2010. IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. *Journal of Medical and Biological Engineering,* 30**,** 23-28.

YANG, Q. & WU, X. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making,* 5**,** 597-604.

YAP, B. W., RANI, K. A., RAHMAN, H. A. A., FONG, S., KHAIRUDIN, Z. & ABDULLAH, N. N. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), 2014. Springer, 13-22.

YEN, S.-J. & LEE, Y.-S. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications,* 36**,** 5718-5727.

YIN, L., GE, Y., XIAO, K., WANG, X. & QUAN, X. 2012. Feature selection for high-dimensional imbalanced data. *Neurocomputing*.

YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J.-F. & HUA, L. 2012. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems,* 36**,** 2431-2448.

ZHANG, D., CHEN, S. & ZHOU, Z.-H. 2008a. Constraint Score: A new filter method for feature selection with pairwise constraints. *Pattern Recognition,* 41**,** 1440-1451.

ZHANG, S., ZHANG, J., ZHU, X., QIN, Y. & ZHANG, C. 2008b. Missing value imputation based on data clustering. *Transactions on computational science I***,** 128-138.

ZHONG, W., CHOW, R. & HE, J. 2012. Clinical charge profiles prediction for patients diagnosed with chronic diseases using Multi-level Support Vector Machine. *Expert Systems with Applications,* 39**,** 1474-1483.

Appendix A : The Datasets

A.1 Malaysian Dataset

A.1.1 Summary of Attributes

No	Attributes	Description	Data Element	Clinical Category	Data Type
1	patientid	Patient Id	DB		ID
2	patientnotifid	Patient Notification Id	DB		ID
3	sdpid	Centre/Hospital Id	Clinical	Admission	ID
4	contactinstitutionname	Centre/Hospital Name	Clinical	Admission	Categorical
5	dateadmission	Date of Admission	Clinical	Admission	Categorical
6	ptdatebirth	Date of birth	Clinical	Demographic	Date
7	yradmit	Year admit	Clinical	Admission	Numerical
8	yrDOB	Year date of birth	Clinical	Demographic	Numerical
9	currentptoutcomeid	Outcome id	Clinical	Clinical Outcome	ID
10	siteid	Centre/Hospital Id	Clinical	Admission	ID
11	patientoutcomeid	Outcome id	Clinical	Clinical Outcome	ID
12	patientfuid	***Unknown*	***Unknown*	***Unknown*	ID
13	ptoutcome30	30 days outcome	Clinical	Clinical Outcome	Categorical
14	dateoutcome	Date of In-hospital outcome	Clinical	Clinical Outcome	Date
15	jpn_dateofdeath	JPN date of death	Non-Clinical	JPN	Date
16	jpn_causeofdeath	JPN cause of death	Non-Clinical	JPN	Categorical
17	jpnmatchingstatus	JPN matching status	Non-Clinical	JPN	Categorical
18	outdate30	Date of 30 days outcome	Clinical	30 days outcome	Date
19	deathdate	Death date	Clinical	Clinical Outcome	Date
20	ptoutcome	In-hospital outcome	Clinical	Clinical Outcome	Categorical
21	yr_outcome	Year of outcome	Clinical	Clinical Outcome	Numerical
22	ptsex	Gender	Clinical	Demographic	Categorical
23	ptrace	Race	Clinical	Demographic	Categorical
24	ptraceothermsian	Other Malaysian race	Clinical	Demographic	Categorical
25	ptraceothermsianspecify	Other specified Malaysian race	Clinical	Demographic	Categorical

26	ptraceforeignspecify	Other specified foreign race	Clinical	Demographic	Categorical
27	ptnationality	Nationality	Clinical	Demographic	Categorical
28	acsstratum	ACS Stratum	Clinical	Clinical Diagnosis	Categorical
29	troponini	Peak Troponin TnI	Clinical	Clinical Investigations and Examinations	Numerical
30	troponinive	Peak Troponin TnI - Positive	Clinical	Clinical Investigations and Examinations	Categorical
31	troponint	Peak Troponin TnT	Clinical	Clinical Investigations and Examinations	Numerical
32	troponintve	Peak Troponin TnT- Positive	Clinical	Clinical Investigations and Examinations	Categorical
33	ultroponini	Reference upper limit for Troponin TnI	Clinical	Clinical Investigations and Examinations	Numerical
34	ultroponint	Reference upper limit for Troponin TnT	Clinical	Clinical Investigations and Examinations	Numerical
35	ptageatnotification	Age at notification	Clinical	Demographic	Numerical
36	smokingstatus	Smoking status	Clinical	Status Before Event - Smoking Status	Categorical
37	statusaspirinuse	Status of aspirin use	Clinical	Status Before Event - Aspirin Used	Categorical
38	cdys	History of dyslipidaemia	Clinical	Status Before Event - Past Medical History	Categorical
39	cdm	History of diabetes	Clinical	Status Before Event - Past Medical History	Categorical
40	chpt	History of hypertension	Clinical	Status Before Event - Past Medical History	Categorical
41	cpremcvd	History of premature cardiovascular disease	Clinical	Status Before Event - Past Medical History	Categorical
42	cmi	History of MI	Clinical	Status Before Event - Past Medical History	Categorical
43	ccap	History of documented cad > 50% stenosis	Clinical	Status Before Event - Past Medical History	Categorical
44	canginamt2wk	History of chronic angina more than 2 weeks ago	Clinical	Status Before Event - Past Medical History	Categorical
45	canginapast2wk	History of chronic angina less than 2 weeks ago	Clinical	Status Before Event - Past Medical History	Categorical
46	cheartfail	History of heart failure	Clinical	Status Before Event - Past Medical History	Categorical
47	clung	History of chronic lung disease	Clinical	Status Before Event - Past Medical History	Categorical
48	crenal	History of renal disease	Clinical	Status Before Event - Past	Categorical

				Medical History	
49	ccerebrovascular	History of cerebrovascular disease	Clinical	Status Before Event - Past Medical History	Categorical
50	cpvascular	History of peripheral vascular disease	Clinical	Status Before Event - Past Medical History	Categorical
51	cnone	None of the stated history disease	Clinical	Status Before Event - Past Medical History	Categorical
52	dateonsetacs	Date on ACS	Clinical	Onset Presentation	Date
53	timeonsetacs	Time onset ACS	Clinical	Onset Presentation	Date
54	timeonsetacsna	(Not Applicable) Time onset ACS	Clinical	Onset Presentation	Categorical
55	dateptpresented	Date presented ACS	Clinical	Onset Presentation	Date
56	timeptpresented	Time presented ACS	Clinical	Onset Presentation	Date
57	timeptpresentedna	(Not Applicable) Time presented ACS	Clinical	Onset Presentation	Categorical
58	transferred	Is patient transferred from another centre	Clinical	Onset Presentation	Categorical
59	anginaepisodeno	Number of distinct episode of angina in past 24hrs	Clinical	Clinical Presentation	Numerical
60	anginaepisodena	(Not Applicable) Number of distinct episode of angina in past 24hrs	Clinical	Clinical Presentation	Categorical
61	heartrate	Heart rate	Clinical	Clinical Presentation	Numerical
62	bpsys	SBP	Clinical	Clinical Presentation	Numerical
63	bpdias	Diastolic BP	Clinical	Clinical Presentation	Numerical
64	height	Height	Clinical	Clinical Presentation	Numerical
65	heightna	(Not Applicable) Height	Clinical	Clinical Presentation	Categorical
66	weight	Weight	Clinical	Clinical Presentation	Categorical
67	weightna	(Not Applicable) Weight	Clinical	Clinical Presentation	Categorical
68	bmi	BMI	Clinical	Clinical Presentation	Numerical
69	waistcircumf	Waist Circumference	Clinical	Clinical Presentation	Numerical
70	waistcircumfna	(Not Applicable) Waist Circumference	Clinical	Clinical Presentation	Categorical
71	hipcircumf	Hip Circumference	Clinical	Clinical Presentation	Numerical
72	hipcircumfna	(Not Applicable) Hip Circumference	Clinical	Clinical Presentation	Categorical
73	whr	WHR	Clinical	Clinical Presentation	Categorical
74	killipclass	Killip class	Clinical	Clinical Diagnosis	Categorical
75	ecgabnormtypestelev1	ECG Abnormalities - ST - segment elevation >=1mm in >=2 contiguous limb leads	Clinical	ECG	Categorical
76	ecgabnormtypestelev2	ECG Abnormalities - ST - segment elevation >=2mm in >=2 contiguous limb leads	Clinical	ECG	Categorical
77	ecgabnormtypestdep	ECG Abnormalities - ST - segment elevation >=0.5mm in >=2 contiguous limb leads	Clinical	ECG	Categorical
78	ecgabnormtypetwave	ECG Abnormalities - T-Wave inversion	Clinical	ECG	Categorical

79	ecgabnormtypebbb	ECG Abnormalities - Bundle Branch Block	Clinical	ECG	Categorical
80	ecgabnormtypenonspecific	ECG Abnormalities - Non specific	Clinical	ECG	Categorical
81	ecgabnormtypenone	ECG Abnormalities - None	Clinical	ECG	Categorical
82	ecgabnormtypenotstated	ECG Abnormalities - Not stated	Clinical	ECG	Categorical
83	ecgabnormlocationil	ECG Abnormalities Location - Inferior leads : II, III, aVF	Clinical	ECG	Categorical
84	ecgabnormlocational	ECG Abnormalities Location - Anterior Leads : V1 and V4	Clinical	ECG	Categorical
85	ecgabnormlocationll	ECG Abnormalities Location - Lateral Leads - I, sVL, v5 and v6	Clinical	ECG	Categorical
86	ecgabnormlocationtp	ECG Abnormalities Location - True Posterior : V1, v2	Clinical	ECG	Categorical
87	ecgabnormlocationrv	ECG Abnormalities Location - Right Ventricle : ST Elevation in Lead V4R	Clinical	ECG	Categorical
88	ecgabnormlocationnone	ECG Abnormalities Location - None	Clinical	ECG	Categorical
89	ecgabnormlocationnotstated	ECG Abnormalities Location - Not Stated	Clinical	ECG	Categorical
90	ckmb	Peak CKMB	Clinical	Clinical Investigations and Examinations	Numerical
91	ulckmb	Upper limit Peak CKMB	Clinical	Clinical Investigations and Examinations	Numerical
92	notdoneckmb	(Not done) Peak CKMB	Clinical	Clinical Investigations and Examinations	Categorical
93	ck	Peak CK	Clinical	Clinical Investigations and Examinations	Numerical
94	ulck	Upper limit Peak CK	Clinical	Clinical Investigations and Examinations	Categorical
95	notdoneck	(Not done) Peak CK	Clinical	Clinical Investigations and Examinations	Categorical
96	notdonetroponint	(Not done) Peak troponin TNT	Clinical	Clinical Investigations and Examinations	Categorical
97	notdonetroponini	(Not done) Peak troponin TN1	Clinical	Clinical Investigations and Examinations	Categorical
98	tc	Total Cholesterol	Clinical	Clinical Investigations and Examinations	Numerical
99	notdonetc	(Not done) Total Cholesterol	Clinical	Clinical Investigations and Examinations	Categorical
100	hdlc	HDL-C	Clinical	Clinical Investigations and Examinations	Numerical
101	notdonehdlc	(Not done) HDL-C	Clinical	Clinical Investigations and Examinations	Categorical
102	ldlc	LDL-C	Clinical	Clinical Investigations and Examinations	Numerical

103	notdoneldlc	(Not done) LDL-C	Clinical	Clinical Investigations and Examinations	Categorical
104	tg	Triglycerides	Clinical	Clinical Investigations and Examinations	Numerical
105	notdonetg	(Not done) Triglycerides	Clinical	Clinical Investigations and Examinations	Categorical
106	fbg	Fasting Blood Glucose	Clinical	Clinical Investigations and Examinations	Numerical
107	notdonefbg	(Not done) Fasting Blood Glucose	Clinical	Clinical Investigations and Examinations	Categorical
108	lvef	Left Ventricular Ejection Fraction	Clinical	Clinical Investigations and Examinations	Numerical
109	notdonelvef	(Not done) Left Ventricular Ejection Fraction	Clinical	Clinical Investigations and Examinations	Categorical
110	timiscorestemi	TIMI Score for STEMI	Clinical	Clinical Diagnosis	Numerical
111	timiscorenstemi	TIMI Score for NSTEMI/UAP	Clinical	Clinical Diagnosis	Numerical
112	fbstatus	Fibrinolytic Therapy status	Clinical	Treatment and Interventions	Categorical
113	fbdrugused	Fibrinolytic Drugs used	Clinical	Treatment and Interventions	Categorical
114	dateivfb	Date intravenous fibrinolytic therapy	Clinical	Treatment and Interventions	Date
115	timeivfb	Time intravenous fibrinolytic therapy	Clinical	Treatment and Interventions	Date
116	doortoneedletime	Door to needle time	Clinical	Treatment and Interventions	Numerical
117	cardiaccath	Patient undergo Cardiac catheterization	Clinical	Treatment and Interventions	Categorical
118	pci	Patient undergo PCI	Clinical	Treatment and Interventions	Categorical
119	pcistemi	Patient received PCI STEMI	Clinical	Treatment and Interventions	Categorical
120	pcistemiurgent	Patient received PCI STEMI - Urgent	Clinical	Treatment and Interventions	Categorical
121	pcistemielective	Patient received PCI STEMI - Elective	Clinical	Treatment and Interventions	Categorical
122	pcinstemi	Patient received PCI NSTEMI	Clinical	Treatment and Interventions	Categorical
123	pcinstemielective	Patient received PCI NSTEMI - elective	Clinical	Treatment and Interventions	Categorical
124	disvesselno	Number of diseased vessels	Clinical	Treatment and Interventions	Numerical
125	lminvolve	Left main stem involvement	Clinical	Treatment and Interventions	Categorical
126	culpritartery	Culprit artery	Clinical	Treatment and Interventions	Categorical
127	date1stangioballoon	Date of First balloon inflation - for urgent PCI	Clinical	Treatment and Interventions	Date
128	time1stangioballoon	Time of First balloon inflation - for urgent PCI	Clinical	Treatment and Interventions	Date
129	doortoballoontime	Door to balloon time - Urgent PCI	Clinical	Treatment and Interventions	Numerical
130	iraprepci	TIMI flow classification pre-PCI	Clinical	Treatment and Interventions	Categorical
131	iraintract	Present of Intra-coronary thrombus	Clinical	Treatment and Interventions	Categorical
132	irapostpci	TIMI flow classification post-PCI	Clinical	Treatment and Interventions	Categorical
133	pcitype	PCI Type	Clinical	Treatment and Interventions	Categorical

134	pcitypestentdirect	Is PCI Type (Stenting) - Direct Stenting	Clinical	Treatment and Interventions	Categorical
135	pcitypestentpredilat	Is PCI Type (Stenting) - Pre dilatation	Clinical	Treatment and Interventions	Categorical
136	pcitypestentbms	Is PCI Type (Stenting) - Drug Eluting	Clinical	Treatment and Interventions	Categorical
137	pcitypestentdes	Is PCI Type (Stenting) - Bare-metal	Clinical	Treatment and Interventions	Categorical
138	cabg	CABG therapy given during admission?	Clinical	Treatment and Interventions	Categorical
139	datecabg	Date of CABG therapy	Clinical	Treatment and Interventions	Date
140	asapre	Aspirin	Clinical	Medical - Pre Admission	Categorical
141	asa	Aspirin	Clinical	Medical - During Admission	Categorical
142	asapost	Aspirin	Clinical	Medical - Post Admission	Categorical
143	adpapre	ADP Antagonist	Clinical	Medical - Pre Admission	Categorical
144	adpa	ADP Antagonist	Clinical	Medical - During Admission	Categorical
145	adpapost	ADP Antagonist	Clinical	Medical - Post Admission	Categorical
146	gpripre	GP receptor inhibitor	Clinical	Medical - Pre Admission	Categorical
147	gpri	GP receptor inhibitor	Clinical	Medical - During Admission	Categorical
148	gpripost	GP receptor inhibitor	Clinical	Medical - Post Admission	Categorical
149	heparinpre	Unfrac Heparin	Clinical	Medical - Pre Admission	Categorical
150	heparin	Unfrac Heparin	Clinical	Medical - During Admission	Categorical
151	heparinpost	Unfrac Heparin	Clinical	Medical - Post Admission	Categorical
152	lmwhpre	LMWH	Clinical	Medical - Pre Admission	Categorical
153	lmwh	LMWH	Clinical	Medical - During Admission	Categorical
154	lmwhpost	LMWH	Clinical	Medical - Post Admission	Categorical
155	bbpre	Beta Blocker	Clinical	Medical - Pre Admission	Categorical
156	bb	Beta Blocker	Clinical	Medical - During Admission	Categorical
157	bbpost	Beta Blocker	Clinical	Medical - Post Admission	Categorical
158	aceipre	ACE Inhibitor	Clinical	Medical - Pre Admission	Categorical
159	acei	ACE Inhibitor	Clinical	Medical - During Admission	Categorical
160	aceipost	ACE Inhibitor	Clinical	Medical - Post Admission	Categorical
161	arb	Angiotensin II Receptor blocker	Clinical	Medical - During Admission	Categorical
162	arbpre	Angiotensin II Receptor blocker	Clinical	Medical - Pre Admission	Categorical
163	arbpost	Angiotensin II Receptor blocker	Clinical	Medical - Post Admission	Categorical
164	statinpre	Statin	Clinical	Medical - Pre Admission	Categorical
165	statin	Statin	Clinical	Medical - During Admission	Categorical
166	statinpost	Statin	Clinical	Medical - Post Admission	Categorical
167	lipidlapre	Other lipid lowering agent	Clinical	Medical - Pre Admission	Categorical
168	lipidla	Other lipid lowering agent	Clinical	Medical - During Admission	Categorical
169	lipidlapost	Other lipid lowering agent	Clinical	Medical - Post Admission	Categorical

170	diureticpre	Diuretics	Clinical	Medical - Pre Admission	Categorical
171	diuretic	Diuretics	Clinical	Medical - During Admission	Categorical
172	diureticpost	Diuretics	Clinical	Medical - Post Admission	Categorical
173	calcantagonistpre	Calcium antagonist	Clinical	Medical - Pre Admission	Categorical
174	calcantagonist	Calcium antagonist	Clinical	Medical - During Admission	Categorical
175	calcantagonistpost	Calcium antagonist	Clinical	Medical - Post Admission	Categorical
176	oralhypoglypre	Oral Hypoglycaemic agent	Clinical	Medical - Pre Admission	Categorical
177	oralhypogly	Oral Hypoglycaemic agent	Clinical	Medical - During Admission	Categorical
178	oralhypoglypost	Oral Hypoglycaemic agent	Clinical	Medical - Post Admission	Categorical
179	insulinpre	Insulin	Clinical	Medical - Pre Admission	Categorical
180	insulin	Insulin	Clinical	Medical - During Admission	Categorical
181	insulinpost	Insulin	Clinical	Medical - Post Admission	Categorical
182	antiarrpre	Anti-Arrhythmic	Clinical	Medical - Pre Admission	Categorical
183	antiarr	Anti-Arrhythmic	Clinical	Medical - During Admission	Categorical
184	antiarrpost	Anti-Arrhythmic	Clinical	Medical - Post Admission	Categorical
185	dayccu	Number of days in CCU	Clinical	Clinical Outcome	Numerical
186	dayicu	Number of days in ICU	Clinical	Clinical Outcome	Numerical
187	totaldaystay	Total number of stay in the hospital	Clinical	Clinical Outcome	Numerical
188	diagatdischarge	Diagnosis at Discharge	Clinical	Clinical Outcome	Categorical
189	bleedingepisodecriteria	Bleeding Complication	Clinical	Clinical Outcome	Categorical
190	zdaygenward	Number of days in general hospital	Clinical	Clinical Outcome	Categorical
191	diff_op	***Unknown*	***Unknown*	***Unknown*	Categorical
192	sdpcode	Centre/Hospital code	Clinical	Admission	Categorical
193	state	State of the admission centre/hospital	Clinical	Admission	Categorical
194	agegp	Age group	Clinical	Demographic	Categorical
195	deathcause	Cause of death	Clinical	Clinical Outcome	Categorical
196	deathcausespecify	Cause of death - specify	Clinical	Clinical Outcome	Text
197	transfercentre	Transfer centre	Clinical	Clinical Outcome	Categorical
198	transfercentrespecify	Specified transfer centre	Clinical	Clinical Outcome	Text
199	yeardeath	Year of death	Clinical	Clinical Outcome	Numerical
200	totaladmday	total admission day	Clinical	Clinical Outcome	Numerical
201	ptoutcome1	Patient outcome	Clinical	Clinical Outcome	Categorical
202	fbstatus_new	Fibrinolytic Therapy status	Clinical	Treatment and Interventions	Categorical
203	dateoutcome30	Date of 30 days outcome	Clinical	Clinical Outcome	Date
204	deathcause30	Cause of death of 30 days outcome	Clinical	Clinical Outcome	Categorical
205	deathcausespecify30	Specified cause of death of 30 days outcome	Clinical	Clinical Outcome	Text

206	transfercentre30	Transfer centre	Clinical	Clinical Outcome	Categorical
207	transfercentrespecify30	Specified transfer centre	Clinical	Clinical Outcome	Text
208	_merge	***Unknown*	***Unknown*	***Unknown*	Categorical
209	ptoutcome30a	30 days outcome	Clinical	Clinical Outcome	Categorical
210	ind	***Unknown*	***Unknown*	***Unknown*	Categorical
211	admission_revised	***Unknown*	***Unknown*	***Unknown*	Numerical
212	admission_string	***Unknown*	***Unknown*	***Unknown*	Categorical
213	DOB_revised	***Unknown*	***Unknown*	***Unknown*	Numerical
214	DOB_string	***Unknown*	***Unknown*	***Unknown*	Numerical
215	age_admit	Age at notification	Clinical	Demographic	Numerical

A.1.2 List of Duplicate Attributes

Set of attributes	Description	Decision/Action Taken
contactinstitutionname, sdpid, siteid, sdpcode	The attributes represents the centre/ hospital that a patient admitted to.	Retained only *sdpid.* However, the reference name of each hospital is kept. The name of each hospital is represented by attribute *contactinstitutionname*.
patientidpatientnotifid	Since it has been decided to have only the first entry of each patient, *patientid* and *patientnotifid* attributes are now considered duplicates as they are both represents unique values.	*Patientid* is used as the unique id for each record in the dataset. Removed *patientnotifid.*
ptageatnotification, age_admit	Both attributes represent the age of a patient on admission.	*Ptageatnotification* is specified in the NCVD data dictionary. Thus, it is assumed that *ptageatnotification* is the true referred attribute for age on admission. Removed *age_admit.*
ptoutcome/ptoutcome1	The attributes hold the in-hospital mortality outcome of a patient.	*ptoutcome* is specified in the NCVD data dictionary. Thus, it is assumed that *ptoutcome* is the true referred attribute in-hospital mortality outcome. Removed *ptoutcome1.*
ptoutcome30, ptoutcome30a	The attributes hold the 30 days hospital mortality outcome of a patient.	Removed *ptoutcome30a.*
dateadmission,Admission_revised	The attributes hold the date of admission of each patient in the dataset.	*dateadmission* is specified in the NCVD data dictionary. Thus, it is assumed that *dateadmission* is the true referred attribute for date of admission. Removed *Admission_revised.*
ptdatebirth, DOB_revised	The attributes hold the date of birth of each patient in the dataset.	*ptdatebirth* is specified in the NCVD data dictionary. Thus, it is assumed that *ptdatebirth* is the true referred attribute for date of birth Removed *DOB_revised.*
totaldaystay, totaladmday	The attributes hold the number of days stay in the hospital of each patient in the database.	*totaldaystay* is specified in the NCVD data dictionary. Thus, it is assumed that *totaldaystay* is the true referred attribute for date of birth Removed *totaladmday*
fb_status/fb_statusnew	The attributes hold the fibrinolytic therapy status of each patient in the database.	Removed fb_statusnew
outcomedate, deathdate.	The attributes hold the outcome date of each patient in the database. Deathdate - The value is in number which cannot be identified on the date. The date	Values in *deathdate* are in numbers which cannot be identified as date. The date of death of a patient is actually the outcome date as we considered death as one

	of death of a patient can be identified by using dateoutcome	of ACS outcome. Removed *deathdate*

A.1.3 List of Database Attributes

Attributes	Descriptions	Decision/Action Taken
patientoutcomeid	The id was generated once the outcome decision is made.	Removed. There is no pattern that will affect the outcome

A.1.4 List of Unknown Attributes

Attributes	Description
currentptoutcomeid	Value is either 3, 5 or blanks. But no specific description of each value representation.
patientfuid	Value is either zero or blanks. But no specific description of each value representation.
outdate30	Each value is unique in numbers and does not represent any pattern
diff_op	Each value is unique in numbers and does not represent any pattern
admission_string	Each value is unique in numbers and does not represent any pattern
DOB_string	Each value is unique in numbers and does not represent any pattern
_merge	Values are either 1, 3 or Blanks
zdaygenward	Probably number of days in general ward.
ptnationality	Probably the nationality of a patient. The value is either 0, 1 or 2. But no specific description of each value representation.

A.1.5 List of Irrelevant Attributes

Attributes	Description
ptraceothermsian	The attribute represents a very specific race yet minority group in Malaysia. The filled out value is very small i.e. only n=29 which might not effect anything towards the outcome. Also, the attribute is very specific towards Malaysian population.
ptraceothermsianspecify	The attribute represents a very specific race yet minority group in Malaysia. The value is captured in text format which may not have any standard. The existing filled out value is very small i.e. only n=59 which might not effect anything towards the outcome. Also, the attribute is very specific towards Malaysian population.
ptraceforeignspecify	The attribute represents the race of foreign patients that are admitted for ACS The value is captured in text format which may not have any standard. The existing filled out value is very small i.e. only n=131 which might not effect anything towards the outcome. Also, the attribute is very specific towards Malaysian population.
ptnationality	Supposedly, the attribute represents the nationality of a patient. The value is either 0, 1, 2, 8888 or 9999. No logical indication can be made from the values and it is not specified in the NCVD data dictionary. Also, since it is a national registry all, patients are mainly Malaysian. Non-Malaysian can be identified by the attribute *ptrace*. This attribute may eventually a duplicate to *ptrace*.
fbdrugused, dateivfb, timeivfb, pcistemi, pcistemiurgent, pcistemielective, pcinstemipcinstemi, disvesselno, lminvolve, culpritartery, date1stangioballoon, time1stangioballoon, doortoballoontime, iraprepci, iraintract, irapostpci, pcitype, pcitypestentdirect, pcitypestentpredilat, pcitypestentbms, pcitypestentdes	These are attributes that describe in details on each therapy or procedure given to a patient. Therapy or procedure is given after doctor has diagnosed the patient. Since the aim of the models is to help doctors or medical practitioners in making diagnosis, therapy or procedure information is not considered as predictors. However, information on type of therapy/procedure a patient received can be the marker to in hospital mortality. It will be beneficial as to use the information to evaluate/analyse the population characteristic. Thus, the information on basic type of therapy/procedure a patient received is remained in the dataset.
transfercentrespecify*, *transferredcentre, transferred	
jpn_dateofdeath, jpn_causeofdeath, jpnmatchingstatus	These are attributes that capture information from NRDM. These have no relation to any of ACS event.
gpripost, heparinpost, lmwhpost	These are attributes that are not captured at all but the attributes exists in the dataset. All values is either 'Missing' or 'blanks'

A.1.6 List of Non-standardized Data Collection Attributes

Attributes
1) troponini
2) troponinive
3) troponint
4) Troponintve
5) Ultroponini
6) ultroponint
7) ckmb
8) ulckmb
9) notdoneckmb
10) ck
11) ulck
12) notdoneck

A.1.7 List of Dependant Missing Attributes

Attributes
1) timeonsetacsna
2) timeptpresentedna
3) heightna
4) weightna
5) waistcircumfna
6) hipcircumfna
7) anginaepisodena

A.1.8 List of New Attributes

No	Attributes	Descriptions	Type	Value
1	*DAYS_ACS_SYMPTOMS_TO_ADMISSION*	Number of days the patient get the symptoms (from the day of admission). Formula: Date of onset of ACS symptoms - Date of admission **NEGATIVE value indicates that ACS symptom before the admission **POSITIVE value indicates that ACS symptom after the admission	Number	0-365

2	*ACS_SYMPTOMS_BEFORE_ ADMISSION*	Indicator whether ACS symptoms were presented before or during admission Formula: If the DAYS_ACS_SYMPTOMS_TO_ADMISSION < 0 and <= -30 then 1 if (DAYS_ACS_SYMPTOMS_TO_ADMISSION > 0 and < 30 days) then 2 else 99	Categoric al	1- True 2- False 99 - NA/Invali d
3	*CNONE*	No past medical history being recorded for the patient. Formula: TRUE - if all past medical history is FALSE FALSE - if any of the past medical history is TRUE Unknown - if all past medical history is Unknown	Categoric al	1- True 2- False 99 - Unknown

A.2 The UK dataset

A.2.1 Summary of Attributes

No	Attribute	Description	Data Source	Data Element	Clinical Category	Data Type
1	ID	Id created for the requested dataset	Link	DB	ID	ID
2	Digest	Pseudonymised ID	SystemOne	DB	ID	ID
3	UNID	Unique ID *** Used for information linkage process*	Link	DB	ID	ID
4	Freq	Indicating multiple records in HES	Link	Clinical	Admission	Numerical
5	STARTAGE	Age at start of episode	HES	Clinical	Demographic	Numerical
6	ETHNOS	Ethnic category	HES	Clinical	Demographic	Categorical
7	SEX	Sex of patient	HES	Clinical	Demographic	Categorical
8	ADMIDATE	Date of admission	HES	Clinical	Admission	Date
9	ADMI_CFL	***Unknown*	***Unknown*	*** Unknown*	*** Unknown*	Numerical
10	ADMIMETH	Method of Admission	HES	Clinical	Admission	Categorical
11	ADMISORC	Source of Admission	HES	Clinical	Admission	Categorical
12	FIRSTREG	First regular day or night admission	HES	Clinical	Admission	Categorical
13	DISDATE	Date of discharge	HES	Clinical	Clinical Outcome	Date
14	DIS_CFL	Discharge date check flag	HES	Clinical	Clinical Outcome	Categorical
15	DISDEST	Destination on discharge	HES	Clinical	Clinical Outcome	Categorical
16	DISMETH	Method of discharge	HES	Clinical	Clinical Outcome	Categorical
17	SPELDUR	Duration of spell	HES	Clinical	Clinical Outcome	Numerical
18	SPELEND	End of spell	HES	Clinical	Clinical Outcome	Categorical
19	EPIORDER	Episode order	HES	Clinical	Clinical Outcome	Categorical
20	DIAG_01	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
21	DIAG_02	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical

22	DIAG_03	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
23	DIAG_04	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
24	DIAG_05	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
25	DIAG_06	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
26	DIAG_07	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
27	DIAG_08	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
28	DIAG_09	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
29	DIAG_10	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
30	DIAG_11	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
31	DIAG_12	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
32	DIAG_13	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
33	DIAG_14	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
34	DIAG_15	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
35	DIAG_16	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
36	DIAG_17	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
37	DIAG_18	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
38	DIAG_19	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
39	DIAG_20	All diagnosis codes	HES	Clinical	Clinical Diagnosis	Categorical
40	OPERTN_01	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
41	OPERTN_02	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical

42	OPERTN_03	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
43	OPERTN_04	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
44	OPERTN_05	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
45	OPERTN_06	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
46	OPERTN_07	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
47	OPERTN_08	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
48	OPERTN_09	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
49	OPERTN_10	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
50	OPERTN_11	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
51	OPERTN_12	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
52	OPERTN_13	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
53	OPERTN_14	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
54	OPERTN_15	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
55	OPERTN_16	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical

56	OPERTN_17	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
57	OPERTN_18	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
58	OPERTN_19	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
59	OPERTN_20	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
60	OPERTN_21	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
61	OPERTN_22	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
62	OPERTN_23	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
63	OPERTN_24	*** Unknown* *** All values are 'NA'*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
64	OPDATE_01	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
65	OPDATE_02	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
66	OPDATE_03	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
67	OPDATE_04	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
68	OPDATE_05	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
69	OPDATE_06	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date

70	OPDATE_07	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
71	OPDATE_08	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
72	OPDATE_09	*** Unknown*	*** Unknown*	Clinical	Treatment and Interventions	Date
73	Column_70	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
74	Column_71	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
75	Column_72	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
76	Column_73	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
77	Column_74	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
78	Column_75	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
79	Column_76	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
80	Column_77	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
81	Column_78	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
82	Column_79	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
83	Column_80	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
84	Column_81	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
85	Column_82	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
86	Column_83	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
87	Column_84	*** Unknown* *All values are 'NA'*	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*
88	Column_85	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical

89	Column_86	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
90	Column_87	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
91	Column_89	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
92	dthdate	*** Unknown* *Used for information linkage process*	Link	DB	*** Unknown*	*** Unknown*
93	dthcode	*** Unknown* *Used for information linkage process*	Link	DB	*** Unknown*	*** Unknown*
94	ADMDATE_COUNT	*** Unknown* *Used for information linkage process*	Link	DB	*** Unknown*	*** Unknown*
95	CVTD_ADMIDATE	*** Unknown* Formatted date from ADMIDATE by Steve mm/dd/yyyy	Link	*** Unknown*	*** Unknown*	Date
96	Ethnicity	Ethnicity	*** Unknown*	Clinical	Demographic	Categorical
97	X107_Gender	Gender	MINAP	Clinical	Demographic	Categorical
98	X306_EventDate	Admission date	MINAP	Clinical	Admission	Date
99	AdmissionDate	Admission date and time	MINAP	Clinical	Admission	Date
100	X107.Gender	Gender	MINAP	Clinical	Demographic	Categorical
101	Ethnic.Group...V83	Ethnic Group	*** Unknown*	Clinical	Demographic	Categorical
102	X201.Admission.Diagnosis	Initial Diagnosis	MINAP	Clinical	Clinical Diagnosis	Categorical
103	Method.of.Admission...V83	Method of Admission	*** Unknown*	Clinical	Admission	Categorical
104	X203.ECG.Determining.Treatment	ECG determining treatment	MINAP	Clinical	ECG	Categorical
105	X204.Where.Aspirin.Given	Where was aspirin/other antiplatelet given	MINAP	Clinical	Status Before Event - Status of Aspirin	Categorical
106	X205.Previous.AMI	Previous AMI	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
107	X206.Previous.Angina	Previous angina	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
108	X207.Hypertension	History of Hypertension	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
109	X208.Hypercholesterolaemia	History of Hypercholesterolaemia	MINAP	Clinical	Status Before Event - Past	Categorical

					Medical History	
110	X209.Peripheral.Vascular.Disease	History of Peripheral Vascular Disease	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
111	X210.Cerebrovascular.Disease	History of Cerebrovascular Disease	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
112	X211.Asthma.or.COPD	History of Asthma or COPD	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
113	X212.Chronic.Renal.Failure	History of Chronic Renal Failure	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
114	X213.Heart.Failure	History of Heart Failure	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
115	X214.Enzymes.Elevated	History of Enzymes Elevated	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
116	X215.Cholesterol	Cholesterol	MINAP	Clinical	Clinical Investigations and Examinations	Numerical
117	X216.Smoking.Status	Smoking Status	MINAP	Clinical	Status Before Event - Smoking Status	Categorical
118	X217.Diabetes	History of Diabetes	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
119	X218.Previous.PCI	Previous PCI	MINAP	Clinical	Status Before Event - Past Treatment	Categorical

120	X219.Previous.CABG	Previous CABG	MINAP	Clinical	Status Before Event - Past Treatment	Categorical
121	X220.Systolic.BP	SBP	MINAP	Clinical	Clinical Presentation	Numerical
122	X221.Heart.Rate	Heart Rate	MINAP	Clinical	Clinical Presentation	Numerical
123	X222.Admitting.Consultant	Type of admitting consultant	MINAP	Clinical	Admission	Categorical
124	X223.Place.ECG.Performed	Place ECG Performed	MINAP	Clinical	ECG	Categorical
125	X224.Beta.Blocker	Beta Blocker	MINAP	Clinical	Medical - Pre Admission	Categorical
126	X225.ACE.I.or.ARB	ACE.I or ARB	MINAP	Clinical	Medical - Pre Admission	Categorical
127	X226.Statin	Statin	MINAP	Clinical	Medical - Pre Admission	Categorical
128	Clopidogrel	Clopidogrel	*** Unknown*	Clinical	Medical - Pre Admission	Categorical
129	X228.Glucose	Serum glucose	MINAP	Clinical	Clinical Presentation	Numerical
130	X229.Height	Height	MINAP	Clinical	Clinical Presentation	Numerical
131	X230.Weight	Weight	MINAP	Clinical	Clinical Presentation	Numerical
132	X231.LVEF	Left ventricular ejection fraction	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
133	X232.Family.History.of.CHD	Family History of CHD	MINAP	Clinical	Status Before Event - Past Medical History	Categorical
134	X233.Cardiological.Care.during.Admission	Cardiological care during admission	MINAP	Clinical	Treatment and Interventions - Cardiac Care	Categorical
135	X234.Creatinine	Creatinine	MINAP	Clinical	Clinical Investigations and Examinations	Numerical
136	X235.Haemoglobin	Haemoglobin	MINAP	Clinical	Clinical Investigations	Numerical

					and Examinations	
137	X236.Site.of.Infarction	Site of Infarction	MINAP	Clinical	ECG	Categorical
138	X237.ECG.QRS.Complex.duration	ECG QRS Complex duration	MINAP	Clinical	ECG	Categorical
139	X238.Thienopyridine.inhibitor.use	Thienopyridine inhibitor use	MINAP	Clinical	Medical - Pre Admission	Categorical
140	X239.Admission.Method	Admission.Method	MINAP	Clinical	Admission	Categorical
141	X240.Patient.location.at.STEMI.onset	Patient location at STEMI onset	MINAP	Clinical	Onset Presentation	Categorical
142	X241.Killip.Class	Killip Class	MINAP	Clinical	Clinical Presentation	Categorical
143	X301.Symptom.Onset	Date Symptom Onset	MINAP	Clinical	Onset Presentation	Date
144	X302.Call.for.Help	Date Call for Help	MINAP	Clinical	Admission	Date
145	X303.Arrival.1st.Responder	Date Arrival.1st.Responder	MINAP	Clinical	Admission	Date
146	X304.Arrival.Ambulance	Date Arrival Ambulance	MINAP	Clinical	Admission	Date
147	X306.Arrival.at.Hospital	Date Arrival at Hospital *** Arrival at the hospital is the same as date of admission*	MINAP	Clinical	Admission	Date
148	X308.Reason.Treatment.not.given	Reason reperfusion treatment not given	MINAP	Clinical	Treatment and Interventions	Categorical
149	X309.Reperfusion.Treatment	Date of Reperfusion Treatment	MINAP	Clinical	Treatment and Interventions	Date
150	X310.Justified.Delay	Delay before treatment	MINAP	Clinical	Treatment and Interventions	Categorical
151	X311.Where.treatment.given	Where was initial reperfusion treatment given?	MINAP	Clinical	Treatment and Interventions	Categorical
152	Who.took.treatment.decision...V7	*** Person who decide on the treatment*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
153	X313.1st.Cardiac.Arrest	Cardiac arrest date/time - FIRST ARREST ONLY	MINAP	Clinical	Onset Presentation	Date
154	X314.Where.cardiac.arrest	Cardiac arrest location	MINAP	Clinical	Onset Presentation	Categorical
155	X315.Presenting.Rhythm	Arrest presenting rhythm	MINAP	Clinical	Onset	Categorical

					Presentation	
156	X316.Outcome.of.arrest	Outcome of arrest	MINAP	Clinical	Onset Presentation	Categorical
157	X317.Admission.Ward	Admission ward	MINAP	Clinical	Admission	Categorical
158	Peak.CK...V7	Peak CK	*** Unknown*	Clinical	Clinical Investigations and Examinations	Numerical
159	X319.Peak.Troponin	Peak Troponin	MINAP	Clinical	Clinical Investigations and Examinations	Numerical
160	X320.Unfractionated.Heparin	Unfractionated heparin	MINAP	Clinical	Medical - During Admission	Categorical
161	X321.Low.molecular.weight.heparin	Low molecular weight heparin	MINAP	Clinical	Medical - During Admission	Categorical
162	X322.Thienopyridene	Thienopyridine platelet inhibitor	MINAP	Clinical	Medical - During Admission	Categorical
163	Other.Oral.Antiplatelet...V7	*** Other Oral Antiplatelet*	*** Unknown*	Clinical	Medical	Categorical
164	X324.IV.2B.3A	IV 2b/3a agent	MINAP	Clinical	Medical - During Admission	Categorical
165	X325.IV.BBlocker	IV beta blocker	MINAP	Clinical	Medical - During Admission	Categorical
166	X327.Calcium.Channel.Blocker	Calcium channel blocker	MINAP	Clinical	Medical - During Admission	Categorical
167	X328.IV.Nitrate	IV nitrate	MINAP	Clinical	Medical - During Admission	Categorical
168	X329.Oral.Nitrate	Oral nitrate	MINAP	Clinical	Medical - During Admission	Categorical
169	X330.Potassium.Channel.Modulator	Potassium Channel Modulator	MINAP	Clinical	Medical - During Admission	Categorical
170	X331.Warfarin	Warfarin	MINAP	Clinical	Medical -	Categorical

171	X332.Angiotensin	Angiotensin	MINAP	Clinical	During Admission Medical -	Categorical
172	X333.Thiazide.Diuretic	Thiazide Diuretic	MINAP	Clinical	During Admission Medical -	Categorical
173	X334.Loop.Diuretic	Loop Diuretic	MINAP	Clinical	During Admission Medical -	Categorical
174	Spironolactone...V7	*** Spironolactone*	*** Unknown*	Clinical	During Admission Medical -	Categorical
175	X336.Thrombolytic.Drug	Thrombolytic Drug	MINAP	Clinical	During Admission Treatment and Interventions	Categorical
176	X337.Troponin.Assay	Troponin Assay	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
177	X338.Fondaparinux	Fondaparinux	MINAP	Clinical	Medical - During Admission	Categorical
178	X339.Initial.Reperfusion.Treatment	Initial Reperfusion Treatment	MINAP	Clinical	Treatment and Interventions	Categorical
179	X340.Additional.Reperfusion.Treatment	Additional Reperfusion Treatment	MINAP	Clinical	Treatment and Interventions	Categorical
180	Was.Reperfusion.Attempted...v6	***** Indicator if any reperfusion attempted*	*** Unknown*	Clinical	Treatment and Interventions	Categorical
181	X341.Inpatient.diabetes.management	Inpatient diabetes management	MINAP	Clinical	Treatment and Interventions	Categorical
182	X342.Diabetic.Therapy	X342.Diabetic.Therapy	MINAP	Clinical	Clinical Outcome - Therapy	Categorical
183	X343.Oral.beta.blocker	Oral beta blocker	MINAP	Clinical	Medical - During Admission	Categorical

184	X344.Aldosterone.antagonist	Aldosterone antagonist	MINAP	Clinical	Medical - During Admission	Categorical
185	X346.Arrival.at.non.interventional.hospital	Arrival at non. intervention hospital	MINAP	Clinical	Admission	Date
186	X347.Assess.at.non.intevention.hospital	Assess at non-intervention hospital	MINAP	Clinical	Admission	Categorical
187	X348.Assess.at.Intervention.Centre	Assess at Intervention Centre	MINAP	Clinical	Admission	Categorical
188	X349.Intended.Reperfusion.Proc	Intended Reperfusion Procedure	MINAP	Clinical	Treatment and Interventions	Categorical
189	X350.Proc.Performed	Procedure Performed	MINAP	Clinical	Treatment and Interventions	Categorical
190	X351.Why.no.Angio	Why no Angio being done?	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
191	X352.Why.no.Intervention	Why no Intervention	MINAP	Clinical	Treatment and Interventions	Categorical
192	X401.Discharge.Date	X401.Discharge.Date	MINAP	Clinical	Clinical Outcome	Date
193	X402.Discharge.Diagnosis	X402.Discharge.Diagnosis	MINAP	Clinical	Clinical Outcome	Categorical
194	X403.Bleeding.Complications	X403.Bleeding.Complications	MINAP	Clinical	Clinical Outcome	Categorical
195	X404.Death.in.Hospital	X404.Death.in.Hospital	MINAP	Clinical	Clinical Outcome	Categorical
196	X405.Discharged.on.Beta.Blocker	Discharged on Beta Blocker	MINAP	Clinical	Medical - Post-Admission	Categorical
197	X406.Discharged.on.ACE.I	Discharged on ACE.I	MINAP	Clinical	Medical - Post-Admission	Categorical
198	X407.Discharged.on.Statin	Discharged on Statin	MINAP	Clinical	Medical - Post-Admission	Categorical
199	X408.Discharged.on.Aspirin	Discharged on Aspirin	MINAP	Clinical	Medical - Post-Admission	Categorical
200	X409.Cardiac.Rehab	X409.Cardiac.Rehab	MINAP	Clinical	Clinical	Categorical

201	X410.Exercise.Test	Exercise Test	MINAP	Clinical	Outcome - Rehab Clinical Investigations and Examinations	Categorical
202	X411.Echocardiography	Echocardiography	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
203	X412.Radionuclide.Study	Radionuclide Study	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
204	X413.Coronary.Angio	Coronary Angio	MINAP	Clinical	Clinical Investigations and Examinations	Categorical
205	X414.Coronary.Intervention	Coronary Intervention	MINAP	Clinical	Treatment and Interventions	Categorical
206	X415.Referral.Date	Referral Date	MINAP	Clinical	Admission	Date
207	X416.Discharge.Destination	Discharge Destination	MINAP	Clinical	Clinical Outcome	Categorical
208	X417.Daycase.Transfer.date	Day case Transfer date	MINAP	Clinical	Clinical Outcome	Date
209	X418.Local.Angio.date	Local Angio date	MINAP	Clinical	Treatment and Interventions	Date
210	X419.Local.Intervention.date	Local Intervention date	MINAP	Clinical	Treatment and Interventions	Date
211	X423.Followed.up	Followed up	MINAP	Clinical	Clinical Outcome	Categorical
212	X424.Reinfarction	Reinfarction	MINAP	Clinical	Onset Presentation	Categorical
213	Discharged.on.Clopidogrel...v7	Clopidogrel (INN)	*** Unknown*	Clinical	Medical - Post-Admission	Categorical
214	X426.Return.to.Referring.Hospital	Return to Referring Hospital	MINAP	Clinical	Clinical Outcome	Categorical

215	X427.Discharged.on.Thieno.Inhibitor	Discharged on Thieno Inhibitor	MINAP	Clinical	Medical - Post-Admission	Categorical
216	X428.Discharged.on.Aldosterone.Antagonist	Discharged on Aldosterone Antagonist	MINAP	Clinical	Medical - Post-Admission	Categorical
217	X429.Interventional.Hospital.Procedure	Interventional Hospital Procedure	MINAP	Clinical	Treatment and Interventions	Categorical
218	X501.Smoking.Cessation.Advice	X501.Smoking.Cessation.Advice	MINAP	Clinical	Clinical Outcome - Advise	Categorical
219	X502.Dietary.Advice	X502.Dietary.Advice	MINAP	Clinical	Clinical Outcome - Advise	Categorical
220	CTH	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Numerical
221	DTN	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Numerical
222	CTN	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Numerical
223	OTH	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Numerical
224	OTN	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Numerical
225	Age.At.Admission	Age at admission	MINAP	Clinical	Demographic	Numerical
226	Apollo...Pseudonymised.103.NHS.Number	Pseudonymised NHS Number	Link	Clinical	ID	ID
227	Hermes...Pseudonymised.101.Hospital.Code	Pseudonymised Hospital Code	Link	Clinical	ID	ID
228	Artemis...Pseudonymised.101.Hospital.Code...102.Hospital.Numbe	Pseudonymised Hospital Number	Link	Clinical	ID	ID
229	Geo...IMDScore	*** Unknown Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical
230	Geo...IMDRank	*** Unknown Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical
231	Geo...HealthScore	*** Unknown Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical
232	Geo...HealthRank	*** Unknown Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical
233	Geo...Easting	*** Unknown Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical

234	Geo...Northing	*** Unknown* *Geographical score*	*** Unknown*	Non-Clinical	Geographical-Score	Numerical
235	validNHS	*** Unknown*	*** Unknown*	*** Unknown*	*** Unknown*	Categorical
236	Sex	Sex of patient	*** Unknown*	Clinical	Demographic	Categorical

A.2.2 List of Duplicate Attributes

Set of attributes	Description	Decision/Action Taken
STARTAGE, Age.At.Admission	Both attributes represent the age of a patient on admission.	Removed *STARTAGE.*
SEX, X107_Gender, X107.Gender, Sex	All attributes represent the gender of a patient	*X107._Gender* is from MINAP Removed *SEX, X107.Gender, Sex*
ETHNOS, Ethnicity, Ethnic.Group...V83	The attributes hold the information about the ethnic group of a patient	***ETHNOS*** is specified in the HES data dictionary. Removed *Ethnicity, Ethnic.Group...V83*
ADMIDATE, X306_EventDate,AdmissionDate, CVTD_ADMIDATE, X306.Arrival.at.Hospital	The attributes hold the information on admission date of a patient	*X306_EventDate* and *AdmissionDate* are from MINAP. *AdmissionDate* is selected because it provides both date and time Removed *X306_EventDate ADMIDATE, CVTD_ADMIDATE, X306.Arrival.at.Hospital*
ADMIMETH, Method.of.Admission...V83, X239.Admission.Method	The attributes hold method of admission	Removed *X239.Admission Method* because it has all blanks values except for 1 record Decided that *ADMINMETH* and *Method of Admission v83* is two different things
DISDATE, X401.Discharge.Date	The attributes hold the date of discharge of a patient	***X401.Discharge.Date***is from MINAP. Since DISDATE is removed, DIS_CFL is also removed since the attribute relates to the existence of DISDATE. Removed *DISDATE*

A.2.3 List of Database Attributes

Attributes
1) Digest
2) Apollo...Pseudonymised.103.NHS.Number
3) Hermes...Pseudonymised.101.Hospital.Code
4) Artemis...Pseudonymised.101.Hospital.Code...102.Hospital.Numbe
5) ADMDATE_COUNT
6) CVTD_ADMIDATE
7) UNID

A.2.4 List of One-value Attributes

Attributes	Descriptions
SPELEND	All values are 'Y'
DIAG_15	All values are 'NA'
DIAG_16	All values are 'NA'
DIAG_17	All values are 'NA'
DIAG_18	All values are 'NA'
DIAG_19	All values are 'NA'
DIAG_20	All values are 'NA'
OPERTN_13	All values are 'NA'
OPERTN_14	All values are 'NA'
OPERTN_15	All values are 'NA'
OPERTN_16	All values are 'NA'
OPERTN_17	All values are 'NA'
OPERTN_18	All values are 'NA'
OPERTN_19	All values are 'NA'
OPERTN_20	All values are 'NA'
OPERTN_21	All values are 'NA'
OPERTN_22	All values are 'NA'
OPERTN_23	All values are 'NA'
OPERTN_24	All values are 'NA'
Column_73	All values are 'NA'
Column_74	All values are 'NA'
Column_75	All values are 'NA'
Column_76	All values are 'NA'
Column_77	All values are 'NA'
Column_78	All values are 'NA'
Column_79	All values are 'NA'
Column_80	All values are 'NA'
Column_81	All values are 'NA'
Column_82	All values are 'NA'
Column_83	All values are 'NA'
Column_84	All values are 'NA'
ADMI_CFL	All values are '0'
validNHS	All values are '1' except for 3 records
KillipClass	All values are 'NA'

A.2.5 List of Unknown Attributes

Attributes	Description
OPERTN_01, OPERTN_02, OPERTN_03, OPERTN_04, OPERTN_05, OPERTN_06, OPERTN_07, OPERTN_08, OPERTN_09, OPERTN_10, OPERTN_11, OPERTN_12	Probably the operation procedure received by the patient. Values are in specific code but details about the attributes are not specified in either HES or MINAP data dictionary.
OPDATE_01, OPDATE_02, OPDATE_03, OPDATE_04, OPDATE_05, OPDATE_06, OPDATE_07, OPDATE_08, OPDATE_09,	Probably the date of operation procedure received by the patient. Values are in date format but details about the attributes are not specified in either HES or MINAP data dictionary.
Column_70, Column_71, Column_72,	Values are in date format but details about the attributes are not specified in either HES or

	MINAP data dictionary.
Column_85, Column_86, Column_87, Column_89	Values are in numeric format but details about the attributes are not specified in either HES or MINAP data dictionary.
Who.took.treatment.decision...V7	Probably the person who made the decision for the treatment. But do not sure for which treatment and details about the attributes are not specified in either HES or MINAP data dictionary.
Other.Oral.Antiplatelet...V7	Probably the medication of other oral antiplatelet given to the patient. But do not sure when the medication is given to the patient and details about the attributes are not specified in either HES or MINAP data dictionary.
Spironolactone...V7	Probably the medication of Spironolactone given to the patient . But do not sure when the medication is given to the patient and details about the attributes are not specified in either HES or MINAP data dictionary.
CTH, DTN, CTN, OTH, OTN	Values are in numeric format but details about the attributes are not specified in either HES or MINAP data dictionary.
Geo...IMDScore, Geo...IMDRank, Geo...HealthScore, Geo...HealthRank, Geo...Easting, Geo...Northing	Probably kind of geographical scores or ranks. But do not sure the meaning of the score or ranks and details about the attributes are not specified in either HES or MINAP data dictionary.

A.2.6 List of Irrelevant Attributes

Attributes	Description
dthdate, dthcode	The attribute was created by the data manager for the purposes of linkage procedure and have no reference to any of the attributes.

A.2.7 List of New Attributes

No	Attributes	Descriptions	Type	Value
1	*ADMISSION_YEAR*	Year of admission for the patient	Categorical	2003 2010
2	*ADMISSION_MONTH*	Month of admission for the patient	Categorical	01-12
3	*ATTEND_NON_INTERVENTIONAL_HOSPITAL*	Indicate that the patient went through non interventional hospital before the interventional FORMULA Based on X346.Arrival.at.non.interventional.hospital. If the date exists, then TRUE else FALSE	Categorical	1-TRUE 2- FALSE
4	*CALL_FOR_HELP*	Indicate that the patient has called for help FORMULA Based on X302.Call.for.Help If the date exists, then TRUE else FALSE	Categorical	1-TRUE 2- FALSE
5	*DEATH_IN_HOSPITAL*	Indicate whether the patient has died in the hospital or not FORMULA Based on X404.Death.in.Hospital. If the (0. No), then 0 else (1. From MI, 2. From complication of treatment, 4. Other cardiac cause, 3. Other non cardiac related cause) 1	Categorical	0 - Not died 1- Died
6	*DAYS_ONSET_SYMPTOMS_TO_ADMISSION*	Number of days the patient get the symptoms before the day of admission. Formula: Date of onset of ACS symptoms - Date of admission **NEGATIVE value indicates that ACS symptom before the admission **POSITIVE value indicates that ACS symptom after the admission if onset of ACS Symptoms is 'blank' then 'BLANK'	Number	0-365
7	*ONSET_SYMPTOMS_BEFORE_ADMISSION*	Indicator whether the ACS symptoms were present before or during admission Formula: If the DAYS_ACS_SYMPTOMS_TO_ADMISSION < 0 then 1 if (DAYS_ACS_SYMPTOMS_TO_ADMISSION > 0) then 2 else 99	Categorical	1- True 2- False 99 -NA/Invalid

A.3 The Mapping of Malaysia and The UK dataset

Category		Attribute Description	The Malaysian Attributes	The Malaysian attributes Value	The UK Attributes	The UK attributes Value
Admission	1	Admission Year	*Yradmit*	[2006 - 2010]	*ADMISSION_YEAR*	[2003 - 2010]
Demographics	2	Age	*ptageatnotification*	Number	*Age.At.Admission*	Number
	3	Gender	*Ptsex*	[Female, Male]	*X107_Gender*	[F,M]
Status before Event - Past Medical History	4	Myocardial Infraction	*Cmi*	[Yes, No, Unknown]	*X205.Previous.AMI*	[Yes, No, Unknown]
	5	Previous Angina	*canginamt2wk & canginapast2wk*	[Yes, No, Unknown]	*X206.Previous.Angina*	[Yes, No, Unknown]
	6	Hypertension	*chpt*	[Yes, No, Unknown]	*X207.Hypertension*	[Yes, No, Unknown]
	7	Peripheral Vascular Disease	*cpvascular*	[Yes, No, Unknown]	*X209.Peripheral.Vascular.Disease*	[Yes, No, Unknown]
	8	Cerebrovascular Disease	*ccerebrovascular*	[Yes, No, Unknown]	*X210.Cerebrovascular.Disease*	[Yes, No, Unknown]
	9	Renal Disease	*crenal*	[Yes, No, Unknown]	*X212.Chronic.Renal.Failure*	[Yes, No, Unknown]
	10	Heart Failure	*cheartfail*	[Yes, No, Unknown]	*X213.Heart.Failure*	[Yes, No, Unknown]
	11	Diabetics	*cdm*	[Yes, No, Unknown]	*X217.Diabetes*	[Yes, No, Unknown]
Status before Event - Smoking Status	12	Smoking status	*smokingstatus*	[Current (any tobacco use within last 30 days), Former (quit >30 days), Never, Unknown]	*X216.Smoking.Status*	[0. Never smoked, 1. Ex smoker, 2. Current smoker, 3. Non smoker - smoking history unknown, 9. Unknown]
Status before Event - Medical Used	13	Beta Blocker	*bbpre*	[Yes, No, Unknown]	*X224.Beta.Blocker*	[1. Yes, 0. No, 9. Unknown]
	14	ACE Inhibitor or Angiotensin II receptor Blocker	*aceipre&arbpre*	[Yes, No, Unknown]	*X225.ACE.I.or.ARB*	[1. Yes, 0. No, 9. Unknown]
	15	Statin	*statinpre*	[Yes, No, Unknown]	*X226.Statin*	[1. Yes, 0. No, 9. Unknown]
Clinical presentation &Examination	16	Heart Rate	*heartrate*	Number	*X221.Heart.Rate*	Number
	17	SBP	*bpsys*	Number	*X220.Systolic.BP*	Number
	18	Height	*height*	Number	*X229.Height*	Number
	19	Weight	*weight*	Number	*X230.Weight*	Number
	20	BMI	*bmi*	Number	*BMI*	Number

ECG	21	ECG Abnormalities Type	*ecgabnormtypestelev1 & ecgabnormtypestelev2 & Ecgabnormtypebbb& Ecgabnormtypestdep& ecgabnormtypetwave*	[TRUE, FALSE]	*X203.ECG.Determining.Treatment*	[1. ST segment elevation, 2. Left bundle branch block, 3. ST segment depression, 4. T wave changes only, 5. Other abnormality, 5. Other acute abnormality , 6. Normal ECG, 9. Unknown]
	22	ECG Abnormalities Location	*ecgabnormlocational& Ecgabnormlocationil& Ecgabnormlocationtp& ecgabnormlocationll*	[TRUE, FALSE]	*X236.Site.of.Infarction*	[1. Anterior, 2. Inferior, 3. Posterior, 4. Lateral, , 5. Indeterminate, 9. Unknown]
Clinical Investigations & Examinations	23	Cholesterol	*tc*	Number	*X215.Cholesterol*	Number
Treatment & Interventions	24	PCI	*pci*	[Yes, No]	*X414.Coronary.Intervention*	[1. Percutaneous coronary intervention]
	25	CABG	*cabg*	[Yes, No]	*X414.Coronary.Intervention*	[2. CABG]
Medical - During Admission	26	Unfrac Heparin	*heparin*	[Yes, No, Unknown]	*X320.Unfractionated.Heparin*	[1. Yes, 0. No, 9. Unknown]
	27	Low molecular weight heparin (LMWH)	*lmwh*	[Yes, No, Unknown]	*X321.Low.molecular.weight.heparin*	[1. Yes, 0. No, 9. Unknown]
	28	Beta Blocker	*bb*	[Yes, No, Unknown]	*X325.IV.Bblocker*	[1. Yes, 0. No, 9. Unknown]
Clinical Outcomes	29	Overnight Stays	*totaldaystay*	Number	*SPELDUR*	Number
	30	Bleeding Complication	*bleedingepisodecriteria*	[Major, Minor, Missing, None, Not Available, Not stated/Inadequately described]	*X403.Bleeding.Complications*	[0. None, 9. Unknown]
	31	Outcome	*ptoutcome*	[Died, Discharge]	*DEATH_IN_HOSPITAL*	[0, 1]

A.4 The Common Dataset

No	Predictors	Type of Predictors	Malaysian Dataset	The UK Dataset
1	Admission year	Categorical	9533(100%) [0%]	3845(100%) [0%]
2	Age	Numerical	59.0 (12.1) [0%]	68.8 (13.4) [0%]
3	Male	Categorical	7225 (75.8%) [0%]	2464 (64.1%) [0%]
4	SBP	Numerical	139.1 (28.7) [1.7%]	147.8 (242.8) [23.1%]
5	Height	Numerical	161.7 (8.3) [45%]	166.1 (65) [70.9%]
6	Weight	Numerical	67.6 (14.1) [38.1%]	78.3 (18.2) [60.6%]
7	Heart rate (beats/mins)	Numerical	83.6 (21.3) [1.7%]	83.7 (34.8) [23.1%]
8	Cholesterol	Numerical	5.31 (1.3) [28%]	11.8 (140.9) [40%]
9	Previous MI	Numerical	1569 (16.5%) [20.8%]	2623 (22.1%) [9.7%]
10	History of heart failure	Categorical	616 (6.5%) [17.2%]	207 (6.5%) [17.3%]
11	History of stroke (cerebrovascular)	Categorical	328 (3.4%) [19.5%]	272 (7.1%) [18.1%]
12	History of peripheral vascular disease	Categorical	74 (1.0%) [20.7%]	195 (5.9%) [13%]
13	History of renal failure	Categorical	586 (7.6%) [19.4%]	159 (5.0%) [18%]
14	History of hypertension	Categorical	5773 (60.6%) [13.8%]	1566 (40.7%) [10.6%]
15	Current smoker	Categorical	3231 (33.9%) [5%]	1009 (26.2%) [12.7%]
16	History of diabetics	Categorical	3964 (41.6%) [17.1%]	567 (14.8%) [8.9%]
17	BB given	Categorical	2269 (27%) [11.9%]	1654(60.5%) [28.9%]
18	Statin given	Categorical	2724 (32.3%) [11.6%]	1993 (72.6%)[28.6%]

A.5 Characteristic of AMIS Model vs. The UK Datasets and Malaysian Datasets.

	(AMIS)-Plus registry	The UK Dataset	Malaysian Dataset
Derivation Population	National Registry (Switzerland)	MINAP Registry (UK)	NCVD Registry (Malaysia)
Years	1997 –2005	2003-2010	2006- 2010
Number of Patients	7520	3846	9533
Source of Patients	54 (out of 106) hospitals treating STEMI in Switzerland	Leeds -selected GP who are using SystemOne, and registered as inpatient and outpatient in the hospitals Leeds	18 hospitals who serve cardiac services in Malaysia
Range of ACS Predictors	UA, NSTEMI, STEMI Age >65 Killip Class >=II SBP Heart Rate Pre-hospital cardiopulmonary resuscitation History of heart failure History of cerebrovascular disease	UA, NSTEMI, STEMI Age Killip Class SBP Heart Rate History of heart failure History of cerebrovascular disease	UA, NSTEMI, STEMI Age SBP Heart Rate History of heart failure History of cerebrovascular disease
In-Hospital Mortality	7.5%	4.8%	7.1%

Appendix B: Results of Methodology Review

B.1 WEKA Classification Algorithms

The grey shaded rows are the 'unsuitable' algorithms.

No	ML Algorithms	Baseline datasets		Common datasets		AMIS	
		Malaysian	The UK	Malaysian	The UK	Malaysian	The UK
1	**BN**	0.778	0.878	0.747	0.726	0.777	0.722
2	**NB**	0.746	0.803	0.754	0.64	0.805	0.656
3	**LG**	0.819	0.809	0.781	0.776	0.797	0.78
4	**MLP**	0.802	0.818	0.718	0.878	0.761	0.756
5	**SVM**	0.500	0.675	0.500	0.500	0.576	0.500
6	**VP**	0.500	0.500	0.500	0.500	0.500	0.500
7	**KNN**	0.615	0.500	0.512	0.495	0.685	0.621
8	**LWL**	0.752	0.849	0.735	0.738	0.762	0.728
9	**CR**	0.500	0.500	0.500	0.500	0.500	0.500
10	**DT**	0.727	0.673	0.781	0.500	0.741	0.500
11	**DTNB**	0.594	0.819	0.435	0.465	0.741	0.500
12	**Jrip**	0.533	0.674	0.521	0.500	0.538	0.500
13	**OneR**	0.505	0.683	0.512	0.500	0.512	0.500
14	**PART**	0.703	0.776	0.761	0.729	0.737	0.500
15	**Ridor**	0.500	0.587	0.509	0.500	0.528	0.500
16	**ZR**	0.500	0.500	0.500	0.500	0.500	0.500

17	**ADT**	0.813	0.868	0.794	0.772	0.794	0.760
18	**BFT**	0.575	0.662	0.589	0.500	0.562	0.500
19	**DS**	0.680	0.672	0.68	0.689	0.603	0.689
20	**FT**	0.721	0.800	0.625	0.500	0.797	0.500
21	**J48**	0.663	0.677	0.615	0.500	0.62	0.500
22	**J48Graft**	0.664	0.677	0.615	0.500	0.62	0.500
23	**LT**	0.732	0.694	0.745	0.753	0.672	0.737
24	**LMT**	0.823	0.851	0.777	0.773	0.797	0.500
25	**NBT**	0.631	0.79	0.752	0.748	0.797	0.745
26	**RF**	0.760	0.767	0.763	0.738	0.714	0.698
27	**RT**	0.638	0.689	0.619	0.648	0.634	0.629
28	**REPT**	0.697	0.670	0.755	0.500	0.768	0.629
29	**SC**	0.589	0.662	0.589	0.500	0.614	0.500

B.2 Missing Values

Algorithm	Malaysian					The UK				
	BD_No_Mssg	BD_5Prct_Mssg	BD_10Prct_Mssg	BD_15Prct_Mssg	BD_20Prct_Mssg	BD_No_Msg	BD_5Prct_Mssg	BD_10Prct_Mssg	BD_15Prct_Mssg	BD_20Prct_Mssg
BN	0.702	0.766	0.765	0.775	0.786	0.717	0.746	0.812	0.816	0.739
NB	0.703	0.779	0.776	0.785	0.796	0.724	0.759	0.824	0.820	0.819
LG	0.707	0.793	0.793	0.799	0.805	0.728	0.753	0.776	0.769	0.771
MLP	0.633	0.715	0.738	0.728	0.732	0.693	0.821	0.801	0.772	0.739
LWL	0.680	0.728	0.719	0.747	0.753	0.712	0.745	0.789	0.779	0.801
DT	0.500	0.635	0.635	0.647	0.647	0.500	0.500	0.730	0.730	0.730
DTNB	0.500	0.449	0.481	0.489	0.596	0.684	0.528	0.637	0.792	0.763
PART	0.626	0.638	0.687	0.639	0.684	0.550	0.633	0.786	0.671	0.738
ADT	0.682	0.778	0.778	0.778	0.778	0.726	0.736	0.784	0.784	0.805
DS	0.618	0.615	0.615	0.615	0.615	0.689	0.689	0.669	0.669	0.669
FT	0.538	0.624	0.659	0.734	0.659	0.741	0.606	0.721	0.678	0.718
LT	0.684	0.751	0.751	0.751	0.751	0.740	0.728	0.780	0.780	0.780
LMT	0.700	0.782	0.782	0.773	0.789	0.733	0.755	0.794	0.798	0.787
NBT	0.701	0.673	0.637	0.604	0.645	0.715	0.744	0.819	0.818	0.818
RF	0.692	0.744	0.757	0.772	0.777	0.673	0.677	0.766	0.762	0.793
RT	0.538	0.577	0.584	0.581	0.580	0.512	0.517	0.603	0.610	0.640
REPT	0.632	0.664	0.677	0.686	0.686	0.649	0.500	0.676	0.676	0.676

Appendix C: Sets of Predictors

C.1 Set of Predictors by Combination of Clinical Categories

Subset of predictors	Malaysian				The UK			
	List of predictors	Training Set	Validation Set	Categorical Predictors	List of predictors	Training Set	Validation Set	Categorical Predictors
CATA1	1) ptsex	6673	2860	18	1) Age.At.Admission	2659	1134	18
	2) ptrace				2) X107_Gender			
	3) ptageatnotification				3) ETHNOS			
	4) smokingstatus				4) X205.Previous.AMI			
	5) statusaspirinuse				5) X206.Previous.Angina			
	6) cdys				6) X207.Hypertension			
	7) cdm				7) X208.Hypercholesterolaemia			
	8) chpt				8) X209.Peripheral.Vascular.Disease			
	9) cpremcvd				9) X210.Cerebrovascular.Disease			
	10) cmi				10) X211.Asthma.or.COPD			
	11) ccap				11) X212.Chronic.Renal.Failure			
	12) canginamt2wk				12) X213.Heart.Failure			
	13) canginapast2wk				13) X217.Diabetes			
	14) cheartfail				14) X232.Family.History.of.CHD			
	15) clung				15) X216.Smoking.Status			
	16) crenal				16) X204.Where.Aspirin.Given			
	17) ccerebrovascular				17) X218.Previous.PCI			
	18) cpvascular				18) X219.Previous.CABG			
	19) CNONE							
CATA2	1) ptsex	6673	2860	34	1) Age.At.Admission	2659	1134	23
	2) ptrace				2) X107_Gender			

3)	ptageatnotificatin	3) ETHNOS
4)	smokingstatus	4) X205.Previous.AMI
5)	statusaspirinuse	5) X206.Previous.Angina
6)	cdys	6) X207.Hypertension
7)	cdm	7) X208.Hypercholesterolaemia
8)	chpt	8) X209.Peripheral.Vascular.Disease
9)	cpremcvd	9) X210.Cerebrovascular.Disease
10)	cmi	10) X211.Asthma.or.COPD
11)	ccap	11) X212.Chronic.Renal.Failure
12)	canginamt2wk	12) X213.Heart.Failure
13)	canginapast2wk	13) X217.Diabetes
14)	cheartfail	14) X232.Family.History.of.CHD
15)	clung	15) X216.Smoking.Status
16)	crenal	16) X204.Where.Aspirin.Given
17)	ccerebrovascular	17) X218.Previous.PCI
18)	cpvascular	18) X219.Previous.CABG
19)	CNONE	19) X224.Beta.Blocker
20)	asapre	20) X225.ACE.I.or.ARB
21)	adpapre	21) X226.Statin
22)	gpripre	22) Clopidogrel
23)	heparinpre	23) X238.Thienopyridine.inhibitor.use
24)	lmwhpre	
25)	bbpre	
26)	aceipre	
27)	arbpre	
28)	statinpre	
29)	lipidlapre	
30)	diureticpre	
31)	calcantagonistpre	
32)	oralhypoglypre	

	33)	insulinpre								
	34)	antiarrpre								
CATA3	1)	ptsex	6673	2860	35	1)	ID	2659	1134	26
	2)	ptrace				2)	Age.At.Admission			
	3)	ptageatnotification				3)	X107_Gender			
	4)	smokingstatus				4)	ETHNOS			
	5)	statusaspirinuse				5)	X205.Previous.AMI			
	6)	cdys				6)	X206.Previous.Angina			
	7)	cdm				7)	X207.Hypertension			
	8)	chpt				8)	X208.Hypercholesterolaemia			
	9)	cpremcvd				9)	X209.Peripheral.Vascular.Disease			
	10)	cmi				10)	X210.Cerebrovascular.Disease			
	11)	ccap				11)	X211.Asthma.or.COPD			
	12)	canginamt2wk				12)	X212.Chronic.Renal.Failure			
	13)	canginapast2wk				13)	X213.Heart.Failure			
	14)	cheartfail				14)	X217.Diabetes			
	15)	clung				15)	X232.Family.History.of.CHD			
	16)	crenal				16)	X216.Smoking.Status			
	17)	ccerebrovascular				17)	X204.Where.Aspirin.Given			
	18)	cpvascular				18)	X218.Previous.PCI			
	19)	CNONE				19)	X219.Previous.CABG			
	20)	ACS_SYMPTOMS_ BEFORE_ADMISSION				20)	X224.Beta.Blocker			
	21)	anginaepisodeno				21)	X225.ACE.I.or.ARB			
	22)	heartrate				22)	X226.Statin			
	23)	bpsys				23)	Clopidogrel			
	24)	bpdias				24)	X238.Thienopyridine.inhibitor.use			
	25)	height				25)	X220.Systolic.BP			
	26)	weight				26)	X221.Heart.Rate			
	27)	waistcircumf				27)	X228.Glucose			
	28)	hipcircumf				28)	X229.Height			

	29)	asapre				29)	X230.Weight			
	30)	adpapre				30)	ONSET_SYMPTOMS_ BEFORE_ADMISSION			
	31)	gpripre				31)	X314.Where.cardiac.arrest			
	32)	heparinpre				32)	X315.Presenting.Rhythm			
	33)	lmwhpre								
	34)	bbpre								
	35)	aceipre								
	36)	arbpre								
	37)	statinpre								
	38)	lipidlapre								
	39)	diureticpre								
	40)	calcantagonistpre								
	41)	oralhypoglypre								
	42)	insulinpre								
	43)	antiarrpre								
CATA4	1)	ptsex	6673	2860	46	1)	ID	2659	1134	30
	2)	ptrace				2)	Age.At.Admission			
	3)	ptageatnotification				3)	X107_Gender			
	4)	smokingstatus				4)	ETHNOS			
	5)	statusaspirinuse				5)	X205.Previous.AMI			
	6)	cdys				6)	X206.Previous.Angina			
	7)	cdm				7)	X207.Hypertension			
	8)	chpt				8)	X208.Hypercholesterolaemia			
	9)	cpremcvd				9)	X209.Peripheral.Vascular.Disease			
	10)	cmi				10)	X210.Cerebrovascular.Disease			
	11)	ccap				11)	X211.Asthma.or.COPD			
	12)	canginamt2wk				12)	X212.Chronic.Renal.Failure			
	13)	canginapast2wk				13)	X213.Heart.Failure			
	14)	cheartfail				14)	X217.Diabetes			
	15)	clung				15)	X232.Family.History.of.CHD			

16)	crenal	16) X216.Smoking.Status
17)	ccerebrovascular	17) X204.Where.Aspirin.Given
18)	cpvascular	18) X218.Previous.PCI
19)	CNONE	19) X219.Previous.CABG
20)	ACS_SYMPTOMS_ BEFORE_ADMISSION	20) X224.Beta.Blocker
21)	anginaepisodeno	21) X225.ACE.I.or.ARB
22)	heartrate	22) X226.Statin
23)	bpsys	23) Clopidogrel
24)	bpdias	24) X238.Thienopyridine.inhibitor.use
25)	height	25) X220.Systolic.BP
26)	weight	26) X221.Heart.Rate
27)	waistcircumf	27) X228.Glucose
28)	hipcircumf	28) X229.Height
29)	ecgabnormtypestelev1	29) X230.Weight
30)	ecgabnormtypestelev2	30) ONSET_SYMPTOMS_ BEFORE_ADMISSION
31)	ecgabnormtypestdep	31) X314.Where.cardiac.arrest
32)	ecgabnormtypetwave	32) X315.Presenting.Rhythm
33)	ecgabnormtypebbb	33) X424.Reinfarction
34)	ecgabnormtypenonspecific	34) X237.ECG.QRS.Complex.duration
35)	ecgabnormlocationil	35) X203.ECG.Determining.Treatment
36)	ecgabnormlocational	36) X236.Site.of.Infarction
37)	ecgabnormlocationll	37) DEATH_IN_HOSPITAL
38)	ecgabnormlocationtp	
39)	ecgabnormlocationrv	
40)	asapre	
41)	adpapre	
42)	gpripre	
43)	heparinpre	
44)	lmwhpre	

	45)	bbpre								
	46)	aceipre								
	47)	arbpre								
	48)	statinpre								
	49)	lipidlapre								
	50)	diureticpre								
	51)	calcantagonistpre								
	52)	oralhypoglypre								
	53)	insulinpre								
	54)	antiarrpre								
CATA5	1)	ptsex	6673	2860	46	1)	ID	2659	1134	33
	2)	ptrace				2)	Age.At.Admission			
	3)	ptageatnotification				3)	X107_Gender			
	4)	smokingstatus				4)	ETHNOS			
	5)	statusaspirinuse				5)	X205.Previous.AMI			
	6)	cdys				6)	X206.Previous.Angina			
	7)	cdm				7)	X207.Hypertension			
	8)	chpt				8)	X208.Hypercholesterolaemia			
	9)	cpremcvd				9)	X209.Peripheral.Vascular.Disease			
	10)	cmi				10)	X210.Cerebrovascular.Disease			
	11)	ccap				11)	X211.Asthma.or.COPD			
	12)	canginamt2wk				12)	X212.Chronic.Renal.Failure			
	13)	canginapast2wk				13)	X213.Heart.Failure			
	14)	cheartfail				14)	X214.Enzymes.Elevated			
	15)	clung				15)	X217.Diabetes			
	16)	crenal				16)	X232.Family.History.of.CHD			
	17)	ccerebrovascular				17)	X216.Smoking.Status			
	18)	cpvascular				18)	X204.Where.Aspirin.Given			
	19)	CNONE				19)	X218.Previous.PCI			
	20)	ACS_SYMPTOMS_ BEFORE_ADMISSION				20)	X219.Previous.CABG			

21)	anginaepisodeno	21) X224.Beta.Blocker
22)	heartrate	22) X225.ACE.I.or.ARB
23)	bpsys	23) X226.Statin
24)	bpdias	24) Clopidogrel
25)	height	25) X238.Thienopyridine.inhibitor.use
26)	weight	26) X220.Systolic.BP
27)	waistcircumf	27) X221.Heart.Rate
28)	hipcircumf	28) X228.Glucose
29)	ecgabnormtypestelev1	29) X229.Height
30)	ecgabnormtypestelev2	30) X230.Weight
31)	ecgabnormtypestdep	31) ONSET_SYMPTOMS_ BEFORE_ADMISSION
32)	ecgabnormtypetwave	32) X314.Where.cardiac.arrest
33)	ecgabnormtypebbb	33) X315.Presenting.Rhythm
34)	ecgabnormtypenonspecific	34) X424.Reinfarction
35)	ecgabnormlocationil	35) X237.ECG.QRS.Complex.duration
36)	ecgabnormlocational	36) X203.ECG.Determining.Treatment
37)	ecgabnormlocationll	37) X236.Site.of.Infarction
38)	ecgabnormlocationtp	38) X215.Cholesterol
39)	ecgabnormlocationrv	39) X231.LVEF
40)	tc	40) X337.Troponin.Assay
41)	hdlc	
42)	ldlc	
43)	tg	
44)	fbg	
45)	lvef	
46)	asapre	
47)	adpapre	
48)	gpripre	
49)	heparinpre	
50)	lmwhpre	

	51) bbpre 52) aceipre 53) arbpre 54) statinpre 55) lipidlapre 56) diureticpre 57) calcantagonistpre 58) oralhypoglypre 59) insulinpre 60) antiarrpre