FORECASTING RISK IN ACUTE MYOCARDIAL INFARCTION

By

Rachel Leigh O'Connell

This thesis is submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy, Macquarie University, July 2011.

Research conducted in association with:

NHMRC Clinical Trials Centre, University of Sydney.

Dedicated to Faye and David O’Connell

TABLE OF CONTENTS

SUMMARY

Coronary heart disease is the most common cause of death worldwide with an estimated 7 million deaths per year. The majority of these deaths are due to acute myocardial infarction (AMI) so the burden of illness and mortality from AMI worldwide is immense. Existing short-term risk assessment strategies in AMI are limited to Western patient populations. In this thesis we have proposed risk models for prediction of mortality after AMI based on the geographically diverse Hirulog and Early Reperfusion or Occlusion (HERO-2) trial. The HERO-2 trial randomised 17 073 patients to either unfractionated heparin or bivalirudin in conjunction with fibrinolytic therapy with streptokinase, for the treatment of ST-segment Elevation MI. Patients were recruited from 46 countries from Europe, North and Latin America and Asia, including Australia, New Zealand and Russia. We have developed a comprehensive risk model to identify significant predictors of 30-day mortality. This model was subsequently simplified to a basic risk index and predictive accuracy was compared. We have also proposed two new methods for directly comparing the calibration and ranking performance of two risk strategies.

The geographical diversity of the HERO-2 trial also provided a unique opportunity to examine international differences in clinical outcomes following AMI. We have undertaken a comprehensive comparison of patient characteristics, treatment and outcomes across 5 pre-specified regions: Western countries, Latin America, Eastern Europe, Russia and Asia. We found that mortality rates were lower in Western countries and that these differences could not be attributed to patient case-mix, treatments or national health and economic statistics.

An important issue in applying findings from randomised clinical trials is the procedure to estimate risk among members of other patient populations. Using the HERO-2 trial we compared methods for updating risk models for AMI. A variety of re-calibration and model revision strategies were compared with a global modeling strategy having a built-in region effect. The relative performance of these methods in the different geographical regions, which vary in sample size, was of primary interest. Model revision was found to only provide a slight improvement in predictive performance over the global model. We concluded that a global model with regional re-calibration is adequate.

We also studied data from 5 additional multinational trials: GUSTO-1, GUSTO-2b, GUSTO-3, ASSENT-2 and ASSENT-3. We further explored the adequacy of applying simple re-calibration to update a model for the context of applying a previously developed model to a new trial. We found that new models do not need to be developed for risk assessment in new trials; prior models with re-calibration will suffice.

STATEMENT OF CANDIDATE

I certify that the work in this thesis entitled "**Forecasting risk in acute myocardial infarction**" has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in this thesis.

Rachel O'Connell (40759350)

November 2010

ACKNOWLEDGEMENTS

I wish to acknowledge access to data from the VIGOUR trials, kindly provided by the trial investigators. This work was partly supported by NHMRC Program Grant 253602.

I owe huge gratitude to the brilliant Professor Malcolm Hudson who was my primary supervisor for most of the duration of this PhD. Malcolm I have learned a tremendous amount from you. You have vastly improved my writing, interpretation and data summarising skills and I have gained substantial new knowledge and understanding of statistical theory and methods from you. These skills and acquired knowledge are priceless and will benefit the remainder of my career immensely. Thank you for your clever ideas and insight.

I would also like to thank Professor Ian Marschner who was involved in this project during the early days and came on board again as my principal supervisor after Malcolm retired. Thanks for allowing me the opportunity all those years ago to carry out the analysis to derive the HERO-2 risk model and teaching me the fundamental skills to develop risk models. More recently thanks for your ideas, feedback and assistance which has been very helpful in finishing off this thesis.

I would also like to thank the NHMRC Clinical Trials Centre, University of Sydney for allowing me the opportunity to undertake this PhD. I am indebted to Professor John Simes who has set up a world class clinical trials research unit. The standard of intellectual contribution and creative thinking among my colleagues is exceptional. This environment has fostered a culture of learning and personal development and has cultivated the research and statistical skills that I am proud to have today. Thanks John for your ideas with the analysis and help with identifying the key messages.

I would also like to acknowledge Professors Hudson, Marschner and Simes for their collaboration in publishing parts of this work as described in full on page viii.

Special thanks to Professor Val Gebski, Adrienne Kirby and Professor Tony Keech for giving me adequate time to devote to this project.

I would also like to thank Rhana Pike for her editorial assistance and help with WORD.

Lastly I would like to thank my family for their enduring support, love and encouragement. It has been a long, arduous road which has taken much perseverance. I could not have achieved this without you.

PUBLICATIONS

The following publications have resulted from work contained in this thesis:

1. O'Connell RL and Hudson HM. Risk of mortality after acute myocardial infarction: Performance of model updating methods for application in different geographical regions. Computational Statistics & Data Analysis 2009; 53(3): 834-46.

2. Simes RJ, O'Connell RL, Aylward PE, Varshavsky S, Diaz R, Wilcox RG, Armstrong PW, Granger CB, French JK, Van de Werf F, Marschner IC, Califf R, White HD; for the HERO-2 Investigators. Unexplained international differences in clinical outcomes after acute myocardial infarction and fibrinolytic therapy: lessons from the Hirulog and Early Reperfusion or Occlusion (HERO)-2 trial. American Heart Journal 2010; 159(6): 988-97.

In both publications I took the lead on all analyses. I also took the lead in writing paper 1 and contributed to the writing of paper 2. For inclusion in this thesis paper 2 has been expanded with the inclusion of additional analyses and rewritten to focus on the statistical aspects of the work.

Other related publications on which I am a coauthor:

3. The Hirulog and Early Reperfusion or Occlusion (HERO)-2 Trial Investigators. Thrombin-specific anticoagulation with bivalirudin versus heparin in patients receiving fibrinolytic therapy for acute myocardial infarction: the HERO-2 randomised trial. Lancet 2001; 358: 1855-63.

4. Edmond JJ, French JK, Stewart RAH, Aylward PA, De Pasquale CG, Williams BF, O'Connell RL, Simes RJ, and White HD, for the HERO-2 Investigators. Frequency of recurrent ST elevation myocardial infarction after fibrinolytic therapy in a different territory as a manifestation of multiple unstable coronary arterial plaques. American Journal of Cardiology 2006; 97(7): 947-51.

5. Edmond JJ, French JK, Aylward PE, Wong CK, Stewart RA, Williams BF, De Pasquale CG, O'Connell RL, Van den Berg K, Van de Werf FJ, Simes RJ, and White HD, for the HERO-2 Investigators. Variations in the use of emergency PCI for the treatment of re-infarction following intravenous fibrinolytic therapy: impact on outcomes in HERO-2. European Heart Journal 2007; 28(12):1418-24.

ACRONYMS

ACS	Acute Coronary Syndrome
AMI	Acute Myocardial Infarction
ANN	Artificial Neural Network
ASSENT-2	Assessment of the Safety and Efficacy of a New Thrombolytic
ASSENT-3	Assessment of the Safety and Efficacy of a New Thrombolytic Regimen
AUC	Area under the curve
BNP	Brain Natriuretic Peptide
CABG	Coronary Artery Bypass Graft
CART	Classification and Regression Trees
CHD	Coronary Heart Disease
CI	Confidence Interval
CT	Computed Tomography
CV	Cross-validation
DALE	Disability Adjusted Life Expectancy
ECG	Electrocardiographic
EPV	Events per variable
GDF-15	Growth Differentiation Factor-15
GISSI	Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico
GNI	Gross National Income
GRACE	Global Registry of Acute Coronary Events
GUSTO-I	Global Utilisation of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries
GUSTO-IIb	Global Use of Strategies to Open Occluded Coronary Arteries in Acute Coronary Syndromes
GUSTO-III	Global Use of Strategies to Open Occluded Coronary Arteries
HERO-2	Hirulog and Early Reperfusion or Occlusion
HF	Heart Failure
HGR	HERO-2 Global Risk
HPI	HERO-2 Prognostic Index
IDI	Integrated Discrimination Improvement
InTIME-II	Intravenous NPA for the Treatment of Infarcting Myocardium Early
IP	Integral of "one minus specificity"
IQR	Interquartile range
IS	Integrated Sensitivity
IV	Intraveneous
LP	Linear Predictor
MAGIC	Magnesium In Coronaries trial
MRI	Magnetic Resonance Imaging
NGRP	Net Gain in Reclassification Proportion
NRI	Net Reclassification Improvement
NSTEMI	Non-ST-segment Elevation MI
NT-ProBNP	N-terminal Pro β-type Natriuretic Peptide
OR	Odds Ratio
PCI	Percutaneous Coronary Intervention
PREDICT	Predicting Risk of Death In Cardiac Disease Tool
PURSUIT	Platelet glycoprotein IIb/IIIa in Unstable angina: Receptor Suppression Using Integrilin (eptifibatide) Therapy
RCT	Randomised Clinical Trial
ROC	Receiver Operating Characteristic
SE	Standard Error
SYSBP	Systolic Blood Pressure
TIMI	Thrombolysis in Myocardial Infarction
t-PA	Tissue Plasminogen Activator
UA	Unstable Angina
VIGOUR	Virtual Coordinating Centre for Global Collaborative Cardiovascular Research

CHAPTER 1: INTRODUCTION

1.1 Risk factors for mortality following acute myocardial infarction

The goal in this thesis is to investigate the benefits of globally estimated mortality prediction models for acute myocardial infarction (AMI), commonly known as heart attack. To date studies examining predictors of short- and long-term mortality following AMI have been based on predominantly Western data [1-11] and we aim to compare these with global models. Emergency room arrival is the most critical point in the clinical course for a patient suspected of having an AMI as urgent therapeutic decisions need to be made which can influence a patient's chances of survival. We focus this investigation on the endpoint 30-day mortality, the most common outcome used in existing risk models in AMI.

Models have been developed using clinical trial data [1, 3-4, 9] and disease registries [7, 10]. Prognostic factors selected for inclusion in risk assessment strategies are very consistent, even including studies using hospital or community registries. Risk stratification tools have also been developed for the general acute coronary syndrome (ACS) population [12-13] and exclusively in patients with unstable angina (UA) or Non-ST-segment Elevation MI (NSTEMI) [14-15]. There is a remarkable homology between patients with different AC syndromes as the same prognostic factors apply across the ACS spectrum (e.g., ST-segment Elevation MI (STEMI) versus UA). Furthermore risk factors for both short- and long-term mortality are very similar.

Age, systolic blood pressure (SYSBP), heart rate and the severity of haemodynamic compromise as measured by Killip class are the key determinants of risk and virtually appear in all AMI models. In a recent model derived from the ASSENT-3 trial data set, age, SYSBP and heart rate accounted for 70% of the relative contribution to the prediction of mortality within the first 3 hours [9]. MI location is the next most influential risk factor and is generally included. The medical history variables prior MI, hypertension and diabetes are also commonly included. The variables which constitute a risk model also depend on the developers desired level of simplicity for that model. Other variables which have appeared in more comprehensive models include: weight, height, time to treatment, history of angina, smoking status, previous coronary artery bypass graft (CABG) surgery and history of cerebrovascular disease. Race and gender have also been found to be independently associated with mortality following AMI.

The information that is available depends on the point in the clinical course that risk assessment is being performed. For example at early presentation baseline information is usually limited to the traditional risk factors described above. Once an electrocardiogram has been performed information on infarct location and other electrocardiographic (ECG) parameters can also be incorporated. The ASSENT-3 baseline risk model identified the ECG parameters ST-segment deviation and QRS score (i.e., index based on Q, R and S waves) as significant predictors of 30-day mortality [9]. QRS was also

identified as a predictor of 1-year survival among 30-day survivors in GUSTO-1 [6]. Selker et al [10] derived a risk model which included data on the amount of ST-segment elevation and Q waves. These ECG indices are difficult to calculate in clinical practice and hence are often not considered for inclusion in risk models. Fortunately the Selker et al model was programmed into an ECG machine with the patients' expected risk and benefit from thrombolytic treatment printed at the top of the ECG output.

Biomarker data has also been shown to provide prognostic information. For example in models developed from the Global Registry of Acute Coronary Events (GRACE) population of ACS patients, creatinine level, an indicator of renal function, was found to be an important predictor of both in-hospital and 6-month post discharge mortality [12, 16]. This variable was also present in a mortality model derived from a Medicare database of elderly AMI patients [7]. This biomarker has received little attention in AMI risk model development as it is not systematically collected in AMI trials, which the majority of existing models are based on, and patients with known renal insufficiency are sometimes excluded from trials. Baseline hemoglobin level has also been shown to be an independent predictor of early mortality in AMI; Giraldez et al [17] combined this laboratory parameter with creatinine clearance to derive a laboratory index which offered additional risk stratification to the TIMI risk score which was developed in patients with STEMI from the InTIME-II trial [3]. Advances in biotechnology in the past decade have demonstrated several other prognostic biomarkers related to cardiovascular outcome, independent of bedside clinical factors. These include cardiac troponin, high sensitivity C-reactive protein, brain natriuretic peptide (BNP), and N-terminal pro β-type natriuretic peptide (NT-ProBNP), which are related to the underlying left ventricular dysfunction or the inflammatory process [18]. The most recent addition is Growth differentiation factor-15 (GDF-15) which is a member of the transforming growth factor-β cytokine superfamily, upregulated and secreted by cardiomyocytes during ischaemia and reperfusion [19]. GDF-15 prognosticated very well in a cohort of STEMI patients receiving fibrinolytic therapy; significant associations with 1-year mortality were demonstrated after adjustment for clinical variables, tropinin T, and NT-ProBNP, differences were already observed within 30 days [19]. However Woo [18] recently highlighted that the immunoradiometric method for GDF-15 assay needs to be further smoothed out for faster reporting in a shorter time.

With the increasing adoption of primary percutaneous coronary intervention (PCI) as a preferable method for achieving reperfusion (i.e., continuous blood flow–and oxygenation through a vascular bed acutely compromised by vasospasm or thrombosis) [20] risk stratification indices have also been developed and validated exclusively in STEMI patients undergoing PCI, for instance the PAMI and CADILLAC risk scores [21-23]. The PAMI risk score can be easily applied at presentation, as it requires information only on age, Killip class, heart rate, diabetes, anterior STEMI and left bundle branch block. However the CADILLAC risk score can only be applied following an angiography performed after PCI as information on triple-vessel disease, final TIMI flow and left ventricular

ejection fraction is required. Therefore the CADILLAC score is mainly relevant to the latter period of hospitalisation which is a limitation as early decisions may be vital for the correct management of patients. Albeit the angiographic parameters included in this score appear to have important additive value to clinical variables as in a study by Lev et al [23] which compared the performance of risk scores using a registry of 974 consecutive patients with STEMI undergoing PCI in an Israel hospital the CADILLAC risk score outperformed, by a large margin, both PAMI and the TIMI risk score developed in patients treated primarily with thrombolytics. Corroborating this Brener et al [24] demonstrated that the angiographic parameter of reperfusion myocardial flow measured shortly after PCI was predictive of 90-day clinical outcomes.

Later in the clinical course information from repeat ECG tests will become available. Further ECG testing also provides an indication of the success of early reperfusion treatment. Brener et al [24] demonstrated that STE resolution is independently associated with the 90-day composite outcome death/congestive heart failure/shock. The use of revascularisation procedures (i.e., PCI or CABG) which have proven benefit can also aid prediction. The occurrence of adverse events such as reinfarction, stroke etc, which are significantly detrimental to a patient's chances of survival, can also be included in risk models. Recently models incorporating such data have been developed in the ASSENT-3 trial population and referred to as dynamic models [9].

We will restrict this investigation of AMI risk stratification tools to models which can be applied at early presentation and which contain traditional risk factors that are easily measured and readily available. This will also enable events occurring early in the clinical course to be used as patients do not have to survive long enough for data which is not immediately available to be measured or integrated into the assessment of risk e.g., biomarkers. This is relevant to both model development and application.

1.2 Uses of risk indexes in AMI

Risk models can assist clinicians in identifying patients at high risk for early mortality following AMI. This can help clinicians to make therapeutic decisions and efficiently allocate treatment recourses. For example a patient identified as at high risk could be considered for treatment with the more expensive and aggressive thrombolytic therapy tissue plasminogen activator (t-PA) instead of the standard treatment streptokinase [2]. Patients identified as being at low risk could be administered streptokinase and may not be considered for further interventions such as revascularisation. This is more cost-effective since high-risk patients derive the largest absolute benefits from treatment with t-PA [2]. This strategy also minimises unnecessary treatment complications. Risk stratification may also help optimise admission and discharge decisions.

Risk models also have other purposes. They provide prognostic information which is valuable information for doctors, patients and their families. A prediction probability or a risk score indicates whether a patient is at high or low risk thereby improving appreciation of the patient's risk level. They

are also useful in understanding biological and causal pathways. Risk indexes are also required in statistical analyses. In clinical trials analyses they provide risk categories needed to examine heterogeneity in treatment effects according to level of risk. In epidemiological studies they are useful in adjusting for baseline risk, e.g., to assess variations in practice patterns or specific therapies between health centres casemix differences need to be adjusted for. Another important purpose is screening for high-risk patients for possible inclusion in or exclusion from clinical trials.

Dynamic risk models aid decision-making later in the clinical course [9]. The enhanced assessment of risk provided by the integration of clinically relevant information unfolding during the hospital stay assists doctors in deciding upon further management, possible early discharge and the level of follow-up care. Since patients must survive to a given time-point for this data to be available these models are based on landmark analysis and provide updated risk forecasts beyond this time-point. Dynamic models are also useful in gaining insights into the disease process and treatment mechanisms.

In general after recovering from AMI, surviving patients remain vulnerable to cardiovascular events such as heart failure (HF), recurrence of angina, reinfarction, arrhythmia, and sudden cardiac death [18]. Risk models which can accurately predict these long-term outcomes are useful to identify high-risk patients needing more intensive follow-up and care.

1.3 Prediction methods

There are a number of methods that can be used for prediction. The most commonly employed are the standard regression approaches, e.g., logistic and Cox regression, for which there is an abundance of statistical advice on how to derive prediction models to maximise model accuracy [25]. These methods are relatively simple and well established with statistical theory that supports and supplements their use (i.e., assessment of adequacy of model fit and verification of model assumptions).

Bayesian model averaging is another approach which involves estimating a number of models and taking the weighted average of predictions from these models to estimate risk [26]. This approach is especially suitable when the training set is small as it accounts for model selection uncertainty. However with large datasets there is little to be gained when used in place of simple regression techniques.

There are also a number of computer intensive classification methods. The most novel of these are classification and regression trees (CART) [27] and artificial neural networks (ANNs) [28]. These approaches are not widely used but have some distinct advantages; they are very flexible because they are non-parametric and allow arbitrary nonlinear relationships as they have no implicit assumptions regarding linearity. CART is an empirical statistical technique that is based on binary recursive partitioning analysis. A series of logical if-then conditions (tree nodes) is produced, and so the output from CART is quite easy for non-statisticians to interpret. One caveat of CART is the seemingly arbitrary selection of split points which may be more data driven than clinically meaningful. ANNs

have shown a lot of promise for diagnostic and prognostic purposes. This complex technique considers the 'chaotic' nature of biologic and pathophysiologic processes and operates by varying the values assigned to nodes and interconnections of a large network modeled in the computer's memory. These values change with experience, and the output of the ANN gradually becomes more accurate as its "experience" increases. ANNs are an excellent complex pattern recognition tool; they have the advantage of detecting complex nonlinear relationships and interactions. The primary disadvantage of ANNs is its "black box" nature; the relationship between the input data and the output diagnosis is obscure. The methods flexibility has led to a lot of recent interest as researcher's anticipated enhanced accuracy. However in a review by Sergeant [29], in large datasets (N>5000) the performance of ANNs and standard regression techniques (i.e., logistic or Cox regression) was equivalent. Additional drawbacks include the computational resources required and the lack of standard software. A major limitation of both CART and ANNs is that they do not allow inference through evaluation of fitted model coefficients.

In AMI Negassa et al [30] derived a simple tree-structured risk stratification model to predict in-hospital mortality for patients undergoing early PCI. The model included three key presenting features: cardiogenic shock, congestive HF, and age. The performance of the model was similar to the Mayo Clinic Risk Score developed to predict risk of major cardiovascular complication following urgent or emergent PCI (discriminatory capacity: 82% versus 80% respectively).

Ennis et al [31] applied a battery of modern, adaptive non-linear learning methods including neural networks and classification trees to the large GUSTO-I database (N=41021). The GUSTO-I trial compared the efficacy of different intravenous thombolytic regimens for the treatment of AMI. Methods were used to predict 30-day mortality and performances were compared to the logistic model of Lee et al [1] also developed in this data set. None of the adaptive non-linear methods outperformed the logistic model. The authors concluded that these adaptive non-linear methods are probably most useful in problems with high signal-to-noise ratio, sometimes occurring in engineering and physical science. In human medical studies, the signal-to-noise ratio is often quite low (as it was in this study) and hence the modern methods may have less to offer.

Despite the potential of these computer intensive methods, logistic regression remains the model of choice for prediction of short-term outcomes and so discussion herein will be limited to this method.

1.4 Risk assessment strategies in AMI

Several types of risk assessment approaches exist in AMI. The first approach is a multivariable logistic model which provides a probability prediction indicating the risk of the event of interest occurring within a specified time period. A linear predictor score is calculated by taking the linear combination of the regression coefficients with the values of the corresponding co-variables. This score is then substituted into the logistic transformation to obtain a probability prediction. Manual computation is laborious and complicated. An example is the GUSTO-I risk model developed in the GUSTO-I trial of

over 40000 patients [1]. This model predicts risk of 30-day mortality and is a complex risk algorithm containing both linear and cubic spline functions and requiring information from 15 input parameters. The availability of a preprogrammed computer or palm pilot is necessary for convenient, timely and correct application.

Another approach is a clinical risk score also referred to as a simple risk index. A risk score is calculated by summing up integer weights assigned to a set of identified risk factors. Patients with a high score are said to be at high-risk and patients with low scores are said to be at low-risk. An example is the already mentioned TIMI Risk Score developed from patients in the InTIME-II trial (N=14114) [3]. The developers suggested that this index could be printed on a laminated card for use at the bedside. On the back of the card would be a table containing the odds ratio for mortality risk relative to average mortality risk associated with each score.

Another similar approach is a nomogram. A nomogram is by definition a graphical calculating device; a two-dimensional diagram designed to allow the approximate graphical computation of a function which uses a coordinate system. A nomogram requires summation of risk points and allows the user to read off the associated risk probability. For ease of use nomograms are often constructed as tables; points are shown alongside specific categories of each prognostic factor with a look up table containing the final risk prediction corresponding to each obtainable overall score. An example is the GUSTO-I nomogram [2] which was developed by simplifying the GUSTO-I model. This nomogram provides the 30-day mortality risk with streptokinase treatment for patients with given characteristics and absolute mortality reduction that would result from substitution with accelerated t-PA treatment.

A disadvantage of the latter two methods is that there is an associated loss of accuracy.

1.5 Global burden of AMI

Acute myocardial infarction, commonly known as a heart attack, occurs when the blood supply to part of the heart is interrupted [32]. This is most commonly due to blockage of a coronary artery following the rupture of a vulnerable atherosclerotic plaque, which is an unstable collection of lipids (like cholesterol) and white blood cells in the wall of an artery. The resulting ischemia (restriction in blood supply) and oxygen shortage, if left untreated for a sufficient period, can cause damage and/or death (infarction) of heart muscle tissue.

Angina pectoris (chest pain) and myocardial infarction are symptoms of and conditions caused by coronary heart disease (CHD; accumulation of atheromatous plaques within the walls of the arteries that supply the myocardium).

CHD is the leading cause of death worldwide and is on the rise due to an increasing incidence in developing and transitional countries [33-34]. This is partly as a result of increasing longevity, urbanisation and life-style changes. In many developed countries the incidence of this disease and specifically AMI is decreasing and case fatality rates have decreased in North America and many western European countries [33, 35]. This decline has been due to improved prevention, diagnosis, and

treatment, in particular reduced cigarette smoking among adults, and lower average levels of blood pressure and blood cholesterol. It is expected that 82% of the future increase in CHD mortality will occur in developing countries.

1.6 Rationale for the thesis

Existing models in AMI have been developed using largely Western patient populations; 80% of subjects in the InTIME-II trial, the basis of the TIMI Risk score, were recruited from Western countries [36] and with the exception of Poland all of the countries that partook in GUSTO-I were Western [37]. Furthermore in the ASSENT-3 trial, the basis for a series of dynamic risk models alluded to earlier, 87% of patients were recruited from Western countries [38]. Given the increasing incidence of CHD, which encompasses MI, it is especially important that risk models are applicable to developing and other non-Western countries. They must encapsulate the correct risk factors with appropriately assigned weights.

We have available data from a large international trial, Hirulog and Early Reperfusion or Occlusion (HERO-2), performed in developed and developing countries [39]. The HERO-2 trial randomised 17073 patients to one of two antithrombotic therapies, unfractionated heparin and bivalirudin in conjunction with fibrinolytic therapy with streptokinase, for the treatment of AMI. The trials' primary endpoint was 30-day mortality. Patients were recruited from 539 hospitals in 46 countries, including Asia, and patients from Eastern Europe, Russia and Latin America were also well represented. Countries were grouped into five prospectively defined geographic regions: Western countries, Latin America, Eastern Europe, Russia and Asia.

This data provides an excellent opportunity to assess the accuracy of existing AMI risk assessment strategies and to propose new models based on a geographically diverse sample. Given the huge sample size and broad spectrum of patients in HERO-2 we anticipate that these new models will have enhanced generalisability and so risk predictions may be improved in non-Western regions. We will ascertain whether the same risk factor effects apply in different regions or do models need to be revised for application to another region. Many findings we present may be applicable to other methods of risk stratification.

1.7 Variations in risk by geographic region – other factors

The HERO-2 clinical trial report [39] revealed large variations in mortality rates and other outcomes by region. If these variations are real, they are of particular concern as they raise a lot of questions concerning the state of health care systems among countries and geographical regions where outcomes were poor. The importance of regional diversity in findings adds to the significance of our evaluation. The structure of the HERO-2 data is hierarchical: patients are treated within hospitals, and hospitals are situated within countries, which in turn were classified into regions. This means there are 3 nested levels contributing to the overall variation.

Differences in patient case-mix factors are the first obvious source of confounding. Patient level factors encompass the traditional risk factors measured in all AMI trials (e.g., demographic characteristics, cardiac history, cardiac risk factors etc) which have proven association with mortality, and unmeasured confounding factors; for example socio-economic status, expressed as education, occupation, income or combinations of these has been shown to be associated with cardiovascular risk and mortality [40-42]. Lifestyle factors have also been shown to affect outcomes following AMI. Regular exercise is associated with a significant reduction in mortality [43-45]. Moderate alcohol consumption has been shown to provide a protective effect against post MI mortality [46], while binge drinking is associated with higher mortality [47]. Diet is also an important factor due to its effects on lipid levels which in turn are related to the incidence of CHD [48]. Ethnic origin has been demonstrated to be independently associated with post-MI mortality with non-caucasians having decreased survival [9]. Genetic factors (e.g., family history of CHD) may also play a part.

Hospital-level covariates including number of beds, volume of cases, presence of on-site catheterisation, percutaneous transluminal coronary angioplasty and open-heart surgical facilities, academic affiliation, for-profit status, total number of physicians, nurses, residents, and other staff may also play a part. The level of utilisation of procedures has been demonstrated to be independently associated with post-MI mortality [49] and the use of evidence-based medicine is a crucial determinant of risk.

Country-level factors such as life-expectancy, economic status and quality of the health care system can also play a role [49].

Another potential source of variation is physician-level factors. The outcomes of patients treated by cardiologists as opposed to by a general practitioner or family physician have been shown to be better [50]. Also the outcomes of patients treated by the same doctor may be correlated. In HERO-2 data to identify individual doctors is not available and so physician-level factors will not be considered.

However in both NSTE ACS and STEMI most of the variation in patient outcomes has been demonstrated to be attributable to patient-level factors (>95%), leaving the remaining small fraction of variation due to physician, hospital and country-level factors [49-51].

Gupta et al [49] conducted a multi-level analysis using the ASSENT-2 trial database to examine the relationship between revascularisation and outcomes following AMI. After adjusting for patient-level characteristics there remained significant variation in 30-day mortality between countries. Further adjustment for country-level life expectancy accounted for much of this variation.

Guigliano et al [52] examined variations in clinical outcomes across four geographical regions, namely Western Europe, North America, Eastern Europe and Latin America in the InTIME-II trial cohort. Despite considering a barrage of hospital-level variables such as those listed above and adjusting for differences in revascularisation rates regional mortality differences were unexplained.

The findings from these two studies are not consistent. Gupta et al [49] listed possible reasons including a smaller gradient in revascularisation rates across the four geographical regions in InTIME-II and differences in analytical approaches employed; Guigliano et al employed standard logistic regression and did not consider country-level factors. Also non-Western patients were more highly represented in the InTIME-II trial as more countries from Eastern Europe participated (11 vs. 3) giving a more geographically diverse sample. Whether variations in outcomes in HERO-2 can be explained will be of key interest; the HERO-2 cohort is the most geographically diverse of any trial to date.

To assess the representativeness of patients in each region enrolled in the HERO-2 study a screening log substudy was conducted. The aim was to ascertain the background patterns of care in the five regions and to compare the trial participants with those initially considered for the HERO-2 trial. The proportion of patients enrolled among those screened may be another factor. Higher risk patients are more often excluded so you would expect higher rates of enrollment to be correlated positively with mortality rates.

1.8 Challenges for models based on randomised trials

Randomised clinical trials (RCT) follow rigorous scientific principles and are not subject to many biases inherent in other designs such as observational or retrospective studies. Therefore they provide a rich source of information which, as required in the conduct of the RCT, is generally accurate and complete. However clinical trial cohorts may not be representative of 'real world' patients due to inclusion and exclusion criteria which restricts the range of patients enrolled in trials. Other factors which make data from clinical trials less realistic are: trial centres and physicians may not be representative of all centres or physicians, patient monitoring is not the same as normal monitoring in practice and health professionals may not act as they would in normal practice [53]. These factors affect the generalisability of clinical trial-derived risk models.

Another difference between clinical trials and normal practice is that information systems and data gathering processes are not the same. This could have implications for the application of risk models. Clinical trials also have the disadvantage that often trial lengths are relatively short which may affect the generalisability of a risk model due to temporal effects.

Disease registries have the advantage that the cohort of patients contained in them are more generalisable since they do not employ strict inclusion and exclusion criteria used in clinical trials and often register consecutive patients admitted to hospital with the disease of interest. However they may be limited in scope or data collection and may not have as stringent data quality, monitoring or endpoint assessment. Since they are non-randomised the formation of inferential statements concerning treatment effects is challenging. Although models developed from the GRACE database, mentioned earlier, appear to have good performance; excellent predictive accuracy was displayed when applied to validation datasets including a subsequent cohort of patients enrolled in GRACE and the GUSTO-IIb trial cohort [12].

The strict inclusion and exclusion criteria imposed in clinical trials is the limitation most relevant to risk models. Clinical trial populations tend to be lower risk than the general population since patients with high-risk features are more often excluded. This reduces the prognostic significance of these variables. Mixed results have been reported for the performance of trial-based risk strategies applied in non-trial cohorts such as disease registries.

In a study by Singh et al [54] which involved a community registry of patients with MI in Olmsted County, Minnesota, the performance of the TIMI risk score was lower compared to in its derivation cohort (discrimination: 73 vs. 78%).

In the study by Lev et al [23], referred to earlier, the performance of various risk scores was compared in a real-life unselected AMI population undergoing PCI using a single-centre registry. The TIMI risk score again did not perform as well compared to in its derivation cohort (discrimination: 72 vs. 78%). However patients selected for PCI generally have fewer exclusion criteria compared with those treated with thrombolytic agents providing a more favorable setting for application of TIMI. The 30-day mortality rate was lower in this study compared to in the TIMI derivation cohort (3.6 vs. 6.7%), which may have hindered the performance of TIMI. The higher efficacy of PCI in establishing vessel patency and brisk epicardial flow compared to thrombolytic therapy would have contributed to this decreased mortality. Even though the GRACE risk score was developed from an ACS registry its performance in this study should be highlighted as it was poor (discrimination = 47%). This raises questions concerning its applicability in exclusively AMI populations.

Rathore et al [55] assessed the performance of the Simple Risk Index, a prediction model encompassing just three predictors age, SYSBP and heart rate, in a community-based cohort of elderly patients (≥ 65 years of age) hospitalised with AMI. The discriminatory capacity of the model was poor compared to in its derivation cohort (71 vs. 78% for 30-day mortality).

Gale et al [56] tested the performance of five risk prediction models derived from trial or registry populations in a national registry of acute hospitals in England and Wales. Models were refitted to derive updated coefficients with all the original variables retained regardless of significance. The discriminatory capacity of all models, including the Simple Risk Index above, was similar and as good as in their original derivation cohorts. This implies that the application of model revision [57] to models developed in trials adequately corrects for bias in regression coefficient effects thus allowing accurate application in non-trial cohorts. The performance of all models was lower in high risk groups.

Finally, using a Canadian ACS registry Yan et al [58] compared the performance of the trial based PURSUIT model to the GRACE model in NSTE patients and found that both models showed similar and good discrimination (84 vs. 83% respectively). The GRACE model demonstrated good calibration, whereas risks predicted by the PURSUIT model were consistently over-estimated as the PURSUIT trial by design selected higher-risk patients.

There was no evidence for the performance of more comprehensive risk models (e.g., GUSTO-I) in STEMI community cohorts without the application of model updating methods. The evidence is limited but it appears that risk scores may not lend themselves to application in real-world cohorts as well as risk models. Their inherent simplicity may be part of the cause – they are derived by approximating regression model coefficients. Risk models are also more amenable to model updating methods such as re-calibration and model revision if needed [57].

1.9 Objectives of the thesis

AIM 1: To develop, using the HERO-2 dataset, new global risk assessment tools for AMI. These will include a comprehensive risk model, a reduced version and a clinical risk score. Of special interest is whether region needs to be included as a risk variable or will differences in mortality rates by region be explained by patient casemix factors. Patient characteristics, treatments and outcomes will be compared across the five geographical regions. Reasons for unexplained regional differences will be further explored using national health and economic statistics.

AIM 2: To compare newly developed indexes to existing Western based strategies. These will include the GUSTO-I risk model and the TIMI risk score, two well known risk indexes developed from large-scale clinical trials. The following questions will be addressed. Can Western based models be successfully (re)-calibrated for application in non-Western patient populations? If so, do the global HERO-2 indexes have a distinct advantage in non-Western patient groups? Conversely, do the Western based models perform better in Western patients? Discriminatory ability will be the main criterion for this comparison.

AIM 3: Clinical risk scores, which are derived by simplifying risk models, are often used in risk stratification. This simplification can involve omitting less important risk factors and categorising continuous variables, the perils of which have been well documented [59]. The cost in predictive accuracy from the resulting loss of information will be assessed using the HERO-2 strategies and recommendations made. The predictive performance of the HERO-2 indexes will be compared and external validity assessed in the GUSTO-I dataset.

AIM 4: Another major issue concerns the use of risk models in new patient populations. Specifically does a new model need to be built from scratch, or is it sufficient to apply simple updating methods to a pre-existing model? Steyerberg et al [57] used the GUSTO-I trial as the basis for a comparison of model updating methods in Western countries. Simple re-calibration methods were found preferable to more extensive model revision methods (re-estimating model coefficients, model extension with more predictors). We extend this evaluation by recalibrating from non-Western patient populations, to determine whether Steyerberg's findings apply in HERO-2, particularly for different geographical regions. Explicitly is there a significant advantage in estimating region-specific coefficients or is a global model with or without the application of re-calibration sufficient? We will also apply model revision with shrinkage of region-specific regression coefficients towards global model coefficients.

AIM 5: To assess the robustness of risk models when applied in other *trial* populations. Recommendations made from AIM 4 will be confirmed for further application in other trials using data from the **V**irtual Coordinating Centre for **G**lobal C**o**llaborative Cardiovasc**u**lar **R**esearch (VIGOUR) project. Specifically what extent of model updating is required when applying a model in another trial? Data from four additional trials will be included: ASSENT-2 [60], ASSENT-3 [38], GUSTO-2b [61] and GUSTO-3 [62]. This will provide data for 42233 additional patients, although only 7% were recruited in non-Western countries; 1381 and 1451 patients in Latin America and Eastern Europe respectively.

1.10 Outline of chapters

Chapter 2 describes all relevant statistical methodology.

Chapter 3 introduces the newly developed HERO-2 risk models and clinical risk score. Simple calibration is used to relate the HERO-2 and TIMI risk scores to outcome to obtain a risk probability for 30-day mortality following the guidelines for (re)-calibration proposed by Steyerberg et al [57]. The performance of the HERO-2 risk indexes are compared with each other and previously developed strategies (GUSTO-I model and TIMI risk score), within geographical regions. New statistical methods that directly compare the predictive performance of risk strategies are proposed. These include a calibration *comparison* measure ("entropy t-test") and a method which allows comparison of error rates in ranking dead-alive pairs between indexes. The consistency of results from these and other methods for comparing the performance of two indices is also considered. The performance of the HERO-2 models and calibrated HERO-2 and TIMI scores is assessed on an external dataset using the GUSTO-I trial data.

Chapter 4 provides a full documentary of regional differences in patient characteristics, patterns of care and clinical outcomes. Patterns of care include administration of protocol treatment, beta-blockers, angiography, revascularisation, duration of hospital stay, CT/MRI scanning and autopsy rates. Outcomes compared include mortality, reinfarction, stroke and severe bleeding. The effect of treatment with bivalirudin versus unfractionated heparin on 30-day mortality and reinfarction outcomes is assessed within regional subgroups. To elucidate reasons for unexplained regional differences in clinical outcomes a multivariable analysis is undertaken which includes adjustment for country-level factors. These factors include national health and economic statistics obtained from external data sources; and rates of coronary revascularisation, durations of hospital stay and percentages of eligible patients enrolled based on the trial data for each country. A sensitivity analysis assessing regional effects will be undertaken using multi-level modeling with hospital and country variables included as random effects. The results of the screening log substudy are presented; patterns of care and patient characteristics are compared in the five regions between trial participants and those screened for possible inclusion in the HERO-2 trial.

Chapter 5 compares the performance of model updating methods in the five geographical regions. Methods are applied to the reduced version of the HERO-2 full model labeled the HERO-2 global risk model. Simple re-calibration (re-estimation of the intercept and slope of the linear predictor within regions) and model revision (re-estimation of all regression coefficients within regions) with and without shrinkage are compared. The relative performance of updating methods is compared across geographical regions. Bootstrap resampling is conducted to obtain optimism-corrected estimates of model performance. The performance of the global model with original weights and Western-specific weights (i.e., derived from the HERO-2 data) when applied in the GUSTO-1 data is compared.

Chapter 6 compares the performance of various risk models across six trials from the VIGOUR project. This will provide information as to the generalisability of risk models when applied in other *trial* populations. Models to be tested will include the HERO-2 (Chapter 3), GUSTO-I [1] and ASSENT-3 [9] full models and a new model also developed in this chapter based on the ASSENT-2 trial data. We will determine whether new models need to be developed from scratch, for a specific trial, or whether recalibrating existing models is adequate. The performance of models in their derivation datasets will be compared with the performance on the same data of models developed in other trials.

The following subsidiary questions will also be contemplated. Do models have an advantage over other models when applied in later related trials given that similar sites participate in serial trials e.g., does the GUSTO-I model display an advantage in the later GUSTO trials? Does one model display a consistent advantage over other models? Data which is independent to all models will aid in this comparison. And finally do models containing simple risk factor information perform as well as models containing difficult to compute indices derived from ECG data?

Further comparisons of the HERO-2 and TIMI risk scores will be conducted. The loss in predictive accuracy from using the HERO-2 risk score in place of the full model in external validation datasets will be studied for consistency with that observed in HERO-2; i.e., it is worth knowing whether the trade-off between simplicity and accuracy seen in HERO-2 extrapolates to other populations.

Chapter 7 will provide a final summary and discussion of the overall findings. Comparisons with other studies will be made as well as recommendations for practical application.

CHAPTER 2: METHODS

2.1 Model building approach

2.1.1 Candidate predictors

We have followed the model building approach proposed by Harrell [25] for developing predictive models as described herein. Choice of candidate covariables should be guided by external findings, prior studies and clinical knowledge. We have therefore only considered common risk factors used in prior studies, ignoring variables with little chance of being predictive or of being measured reliably. This ensures less over-fitting and greater generalisability.

2.1.2 Missing data

In some circumstances, deletion of cases with missing predictors may cause bias and increased variance in regression parameter estimates. Harrell [25] advocates imputing missing values as opposed to complete case analysis which should only be performed if the dataset is very large and offers no advantage apart from saving the analyst time. Harrell's [25] rough guideline specifies that if the proportion of observations with any variables missing is less than 5% simple imputation methods are adequate. This involves imputing missings with the median nonmissing value; for categorical predictors the most frequent category can be used. Harrell [25] states that in this case it does not matter how missings are imputed or whether the variances of regression coefficient estimates are adjusted for having imputed.

If the proportion of cases with missing data is between 5 and 15% the guidelines are as follows. If a predictor is unrelated to all of the other predictors, imputations can be done the same as above. If the predictor is correlated with other predictors, develop a customised model to predict the predictor from all other predictors. Then impute missings with predicted values. For categorical variables, classification trees, binary or polytomous logistic regression are good methods for developing customised imputation models. For continuous variables, ordinary regression can be used. This is referred to as the "best guess" imputation method which fills in missings with predicted expected values from using a multivariable imputation model based on non-missing data. Variances may need to be adjusted for imputation. This is conditional mean imputation which is probably adequate for this level of missing data, but multiple random-draw imputation may be preferred.

Because new risk models are being developed in HERO-2 the approach for dealing with missing data is crucial. Fortunately the frequency of subjects with missing data was very low (<5%) so model parameters could be estimated with sufficient certainty and multiple imputation was not warranted. We chose to preserve the whole sample and apply simple imputation for variables with less than 1% of values missing. For variables with greater than 1% of values missing the appropriate regression method was applied to predict the predictor from all other predictors where sensible (i.e., conditional mean imputation). Otherwise reduced models derived using backward selection was used to impute

missing values. Multivariable imputation models were derived in subjects with non-missing data for the predictors included in these models. The low rate of missing data also ensures that single imputation methods are sufficient to obtain reasonable predictions.

In the case where the covariate X_j needs to be imputed for some subjects based on other variables that themselves were missing on the same subjects missing X_j the following algorithm was applied. First, for the variables considered more important all missing values were initialised to medians (modes for categorical variables). Then for the variables considered less important missing values were imputed using the derived prediction equation. And finally imputed values for the variables considered more important were updated using their respective prediction equations and imputed values for the variables considered less important obtained in the previous step.

Refer to Chapter 3 for more information on missing data rates in HERO-2. In Chapters 3 and 5 the GUSTO-I dataset was used for external validation of the HERO-2 risk strategies. For simplicity simple imputation was applied for all predictors with missing data in GUSTO-I. This imputed dataset was later copied across for use in the cross-trial comparisons (Chapter 6). However for the other VIGOUR datasets missing data was imputed according to the same rules applied in HERO-2.

Fortunately missing data rates for all the key predictor variables were very low across all trials. Refer to Chapter 6 for more information on missing data rates in VIGOUR.

Harrell [25] states that if there are missing Y (i.e., the dependent variable) values on a small fraction of the subjects but Y can be reliably substituted by a surrogate response, use the surrogate to replace the response. In this setting the variable discharge status was useful for that purpose as most patients who were discharged alive survived until day 30. In any case the frequency of subjects with missing data for 30-day mortality status was very low (< 0.5%) across all trial datasets. Patients were assumed alive at 30 days unless there was information to suggest otherwise except in Chapters 3 and 5 where GUSTO-I patients with missing mortality status were deleted.

2.1.3 Functional form

We examined the relationship between continuous variables and mortality using generalised additive models since prior studies have unveiled non-linear effects for some variables particularly height, SYSBP, heart rate and time from symptom onset [1]. Generalised additive models, a flexible nonparametric model-fitting approach involving piecewise cubic polynomials, enable the user to determine visually from a plot the appropriate functional form whilst adjusting for other covariates [63-64]. Because of the large sample size of HERO-2 and number of events additional degrees of freedom could be devoted to continuous predictors whose functional form required additional terms.

2.1.4 Interactions

Because a prior study uncovered a significant interaction between Killip class and age [1] testing of this interaction was pre-specified. To assess whether other interactions were necessary as a first step global interaction tests were conducted for each variable before pursuing component interactions. This

involves, for each predictor separately, testing simultaneously the joint importance of all interactions involving that predictor. Significant interactions can manifest due to lack of model fit or multiple testing (type I error) and hence were only retained if considered biologically plausible. Also interactions which were statistically significant but not perceived as clinically significant were discarded. Only interactions between variables which were significantly associated with 30-day mortality in multivariable analysis were considered.

2.1.5 Overfitting considerations

The number of terms fitted or tested (counting non-linear and cross-product terms) in the modeling process relative to the number of events in the training dataset is an important consideration. For example if too many predictors are fitted the user may end up fitting not only real trends in the dataset but also peculiarities which exist in that particular dataset resulting in poor performance when the model is applied to new patients. Harrell [25] advocates that the predictor degrees of freedom should be no more than m/10, where for logistic regression m is the number of patients in the less frequent outcome category. If the predictor degrees of freedom is larger than this data reduction should be applied until the number of remaining free variables needing regression coefficients is tolerable. In HERO-2 there were 1850 patient deaths, so up to 185 predictor degrees of freedom was permitted for model development using the whole dataset. This was adequate for the approach taken to develop the HERO-2 full model which commenced with 19 candidate variables.

2.1.6 Variable selection

Harrell [25] advocates that the variables that are included in a prediction model should be pre-specified as opposed to using stepwise procedures. The main arguments for this are: stepwise methods provide regression coefficients that are biased high in absolute value, yield standard errors biased low and confidence intervals that are falsely narrow. There are also severe multiple comparison problems leading to p-values that are too small. For ease of application we desired a parsimonious model and therefore only selected significant predictors for inclusion in the final model. We used the standard significance level for testing of hypotheses (α=0.05). Furthermore given the high power afforded by the large HERO-2 dataset, we were only interested in retaining variables which demonstrated a significant effect on the outcome. The caveat to this is that risk factors which are rare may be overlooked because of a lack of power. However the estimation of their effect may be unreliable anyway.

Steyerberg [65] advocates that if the events per variable (EPV) ratio for a given dataset is very high (>50) application of stepwise procedures should still produce a valid model; the selection bias should be negligible. The HERO-2 dataset fulfills this criterion.

Backward elimination is the recommended stepwise procedure as it usually performs better than forward stepwise methods, especially when collinearity is present; with the backward approach correlated variables may remain in the model, while none of them might enter the model with a

forward approach [65]. Also the modeller is forced to examine the full model fit, which is the only fit providing accurate standard errors, error mean square or explained deviance, and P-values. This also enables the effects of all candidate predictors to be judged simultaneously.

2.1.7 Derivation dataset

Harrell [25] advocates using the entire sample for model development as data are too precious to waste. To obtain an accurate assessment of predictive performance models need to be tested on a separate dataset or bias-corrected estimates on the training sample need to be calculated using bootstrapping. However with large sample sizes and event numbers (e.g., EPV ≥40), optimism in performance measures has been found to be small and apparent estimates of model performance are attractive because of their stability [66] and so bootstrapping was not indicated for our final model.

For models fitted in subsections of the data (i.e., regions) bias-correction of model performance estimates was warranted and so bootstrapping was conducted. The specific details of this analysis are provided in Chapter 5. An explanation of the bootstrapping methodology is provided below (Section 2.3).

2.1.8 Simplifying the final model

We also endeavored to approximate our full model to derive a more parsimonious version which requires less information to apply. Model simplification involves a bias-variance tradeoff. The full model requires predicted values to be conditional on all of the predictors, which can increase the variance of the predictions. A predicted value will have a higher variance if it is for a subject in a minority group for an included covariate compared to a model that excluded this covariate. Omission of relevant predictors introduces bias into predictions.

When approximating full models one stops deleting variables when deleting any further variable would make the approximation inadequate. Harrell [25] suggests when the R^2 for predictions from the reduced model against the original model drops below 95%. We applied this rule to derive the HERO-2 reduced model also referred to as the HERO-2 global risk model.

A succinct summary of the exact steps followed to derive the HERO-2 risk strategies is provided in Chapter 3.

2.2 Predictive performance

To assess predictive performance we used methods currently in wide use in medical journals as recommended by Steyerberg [65]. There are two aspects to consider: the discriminatory ability (a model's ability to correctly rank patients), and calibration (how closely predicted probabilities agree with observed probabilities).

Discriminatory ability is the primary requirement of a prediction model to identify a high risk group, or to perform covariate adjustment of a randomised controlled trial. Calibration is the primary requirement for informing patients and medical decision making.

To assess discrimination we used *c* statistics (identical to the area under the receiver operating characteristic (ROC) curve (AUC), ranging from 0 to 1) [67]. For calibration we used calibration plots, assessed calibration-in-the-large and the calibration slope and lastly performed an overall test of miscalibration.

In calibration plots subjects are grouped by similar probabilities (quantiles) and the mean predicted probability is compared to the mean observed outcome. The choice of quantiles is important for the visual impression of calibration; if small groups are plotted, the variability will be large. A better discriminating model has more spread between such quantiles than a poorly discriminating model. Deciles of predicted risk were used to group subjects. Where deciles were not appropriate because of the discrete nature of the risk index or because too large a range of (high) predicted risks were included in the upper decile, categories were chosen so that total expected event counts exceeded 50. We also included in these plots smooth nonparametric calibration curves fitted to the data using lowess to show the correspondence between predicted probabilities and observed mortality for the entire range of the risk distribution. The range of the dotted/dashed line for these curves corresponds to the limit of the data. Because of low numbers of events at the low end of predictions we were mindful not to over-interpret the extremes of the data. Below each calibration plot histograms were also shown depicting the relative frequency distribution of predicted probabilities. Since calibration plots were constructed with logarithmic axes this meant that the plotted width of histogram bars were larger for lower probabilities and decreased as risk increased along the x-axis. This gave the false visual impression that there is more data at the low-risk end of the distribution than in actuality. To remedy this we adjusted the heights of the bars so that the area within each bar is indicative of the frequency of observations in that interval relative to other intervals. Calibration plots were constructed using a modified version of the function val.prob written by Harrell [68] and included in the Design library of S-PLUS version 8.0 for Windows.

Calibration-in-the-large is the agreement between the average predicted risk and the average observed risk [69]. In model development this correspondence is guaranteed by the intercept in a generalised linear model and remains at internal validation with bootstrapping. However in external validation this correspondence may be less. The difference in average predicted and observed risk was tested for statistical significance by fitting logistic regression models for each risk model with the linear predictor included as an offset variable (i.e., the regression coefficient of the linear predictor was fixed at unity). The intercept was the only free parameter; it would be zero with perfect calibration-in-the-large.

The calibration slope, as originally proposed by Cox [70], assesses the average strength of the predictor effects. It is the regression coefficient β in a logistic model with the linear predictor (i.e., the combination of regression coefficients from the model and the predictor values in the new data) included as the only covariate: observed mortality logit = $\alpha + \beta$ linear predictor. A slope less than 1 indicates that shrinkage is required as the model is providing too extreme predictions for low- and

high-risk subjects. Conversely a slope greater than 1 indicates inflation is required as predictions are too high and too low for the lowest- and highest-risk subjects respectively. At internal validation the calibration slope indicates how much we need to reduce or increase the effects of predictors on average to make the model well calibrated for new patients from the underlying population. Hence it can be used as a shrinkage or inflation factor for future use. At external validation, the calibration slope reflects the combined effects of two issues: overfitting on the development data and true differences in effects of predictors.

The overall miscalibration test compares the deviance difference of a model with α and β above included as free parameters and a model with α = 0 and β = 1. The test has two degrees of freedom. The miscalibration test can pick up common patterns of miscalibration, i.e. systematic differences between the new data and the model development data, and overfitting of the effects of predictors.

At external validation a smaller range in predictions may arise from a narrow selection of patients (homogeneous case-mix). A drop in discriminatory ability compared with the development setting can hence be explained by overfitting (calibration also poor), a more homogeneous case-mix, or differential effects in predictors.

We also studied overall measures of performance such as Nagelkerke's R^2 [71] and the Brier score [72]. These measures incorporate both calibration and discrimination aspects. Nagelkerke's R^2 provides a measure of explained variation of a fitted model and is calculated from the log-likelihood. This measure was proposed as a modification to an earlier more traditional definition in order to derive an R^2 which ranges from 0 to 1. This is achieved by dividing by its maximum attainable value as follows:

$$R^2 = 1\text{-}\exp[-2/n\{l(\hat{\beta})\text{-}l(0)\}],$$

$$\max(R^2) = 1\text{-}\exp\{2n^{-1}l(0)\},$$

and

$$\overline{R}^2 = R^2/\max(R^2),$$

where $l(\hat{\beta})$=log $L(\hat{\beta})$ and $l(0)$=log $L(0)$ denote the log likelihoods of the fitted and the 'null' model respectively. The fitted model uses the risk model linear predictor as an offset, with no constant term.

The Brier score uses a quadratic scoring rule and is calculated as $\sum(y_i - p_i)^2/n$, where y denotes the observed outcome and p the prediction for subject i in the dataset of n subjects. For sensible models, the Brier score ranges from 0 (perfect) to 0.25 (worthless). A disadvantage of the Brier score is that the interpretation depends on the outcome. The Brier score is lower, at a lower outcome frequency, and so scores for the various risk strategies will only be compared for common patient populations (i.e., within regions).

Studying discriminatory ability and calibration is often more meaningful than an overall measure such as R^2 or Brier score when we want to appreciate the quality of model predictions for individuals.

Calibration is only considered at the region-level as opposed to country- or hospital-level adjustment which is beyond the scope of this thesis.Further details for the application of these methods are provided in the relevant chapter. Additional methods which are specific to each chapter are also described later.

2.3 Internal and external validation

In internal validation the validity of the prediction model is assessed for the setting where the development data originated from. There are numerous techniques for assessing internal validity. The techniques utilised in this work include apparent validation and bootstrapping which are now described.

With apparent validation, model performance is assessed directly in the sample where it was derived. This procedure gives optimistic but stable estimates of performance. However, as mentioned above, (Section 2.1.7) optimism is small in large samples.

Bootstrapping mimics the process of sampling from the underlying population [73]. Samples, of the same size as the original sample, are drawn with replacement from the original sample. For model validation, 100-200 bootstraps is often sufficient to obtain stable estimates; 200 was used in this thesis. The procedure involves developing a prediction model in each bootstrap sample and subsequently evaluating this model (e.g., using *c* statistics), separately, in both the bootstrap and original samples. The difference in performance indicates the optimism. The average optimism is subtracted from the apparent performance of the original model in the original sample. Bootstrapping has been demonstrated to perform better than other methods such as split-sample and cross-validation; optimism-corrected performance estimates are more stable, since samples of size N are used to both develop and test the model [66].

In external validation the "transportability"/"generalisability" of the prediction model to populations that are "plausibly related" is assessed. The types of external validity considered in this work are now described. With temporal validation the validity of a model is assessed in more recently treated patients. For geographic validation, a predictive model is evaluated in patients from other sites. Spectrum transportability refers to testing in patients who are, on average, more (or less) advanced in their disease process, or who have a somewhat different disease. An example of this is validating a model developed in a randomised trial in a broader, less-selected sample such as in patients from a disease registry.

There are many possible reasons for poor validity. Providing the quality of the prediction model as developed for the developmental setting is good, a model may miscalibrate due to systematic differences between the development and validation samples. Differences may involve encoding of predictors, missed predictors with different distributions, and truly differential effects [65].

2.4 Updating prediction models for a new setting

To apply a model in a new setting the model may need to be updated for local and/or contemporary circumstances. The updating strategies employed in this work are briefly explained here.

Updating the intercept in a logistic regression model corrects calibration-in-the-large; i.e., the correspondence between the mean observed mortality in the new setting and the mean of the predicted outcome. This intercept adjustment corrects for differences between settings in other aspects not captured by the predictors.

Logistic calibration involves fitting a model in the new setting with the linear predictor based on the previously developed model included as the only covariate. This calibration model has two free parameters, intercept α and calibration slope $\beta_{overall}$. The updated model will have a new intercept α and new regression coefficients based on multiplication of the original coefficients with $\beta_{overall}$.

Model revision involves re-estimation of all regression coefficients in the new setting.

Model revision with additional *shrinkage* involves re-estimation of regression coefficients in the new setting with the additional application of shrinkage of revised coefficients towards their original values.

The first two strategies described are *re-calibration* methods, the latter two are model revision.

Specific details for the application of these methods are provided in Chapter 5.

2.5 Software

All analyses were performed using SAS v. 9.1 (SAS Institute Inc.), S-PLUS v. 7-8 (Insightful Corp.) and Stata v. 9.1 (StataCorp LP) statistical software.

CHAPTER 3: RISK MODELS FOR ACUTE MYOCARDIAL INFARCTION BASED ON A GEOGRAPHICALLY DIVERSE TRIAL POPULATION

3.1 Introduction

The validity of risk stratification models beyond the populations in which they were formed is an empirical question of great interest. Results of the large international trial, HERO-2, performed in developed and developing countries, allow the accuracy of existing AMI indexes that predict mortality risk of individuals following MI to be studied and their performance compared with new indexes developed in this trial. Risk stratification is often based on simplified clinical indexes, but there remains some divergence between clinical and statistical commentaries about appropriate variable coding and model selection methods. Derived indexes resulting from model selection can omit relevant prognostic indicators in forming a minimal set of risk factors. This can introduce bias in risk estimates [74]. Disadvantages of categorising continuous variables have been well documented in the literature [75]. In the research area of model development for the purpose of risk prediction there are several pertinent issues. First, categorisation results in a substantial loss of statistical power to detect relations with outcome [59] and second, implies the false assumption of a piecewise flat relationship. A third issue is how to determine cutpoints. Choice may depend on clinical considerations, prevalence in the population, results in other published studies or on outcome-orientated statistical techniques that select the cutpoint that results in the most significant relation with outcome. These 'Optimal' cutpoint methods have serious drawbacks such as inflated type I error rates due to multiple testing [76], over-estimated parameter estimates [76-77] and confidence intervals which are too narrow [78]. Methods have been proposed which correct for these limitations [76-78], but they are cumbersome. The dependence of these modeling approaches on the sample composition, outliers and other variables included in the model should also be contemplated. Therefore the effectiveness of models simplified by model reduction and simple risk scores containing only categorical variables with integer weights will also be studied.

To reiterate, the HERO-2 trial was a large multi-national randomised trial of two antithrombotic therapies, unfractionated heparin and bivalirudin in conjunction with fibrinolytic therapy with streptokinase, for the treatment of AMI. The primary endpoint was 30-day mortality. In this chapter risk strategies developed from the HERO-2 data (i.e., a multivariable model with continuous variables, a reduced version of this model and a simple risk score) and previously developed TIMI Risk Score [3] and GUSTO-I [1] risk model which were also developed using data from large RCTs are compared.

HERO-2 recruited patients from 539 hospitals in 46 countries, including Asia, and patients from Australia and New Zealand, Eastern Europe, Russia and Latin America were also well represented. We will assess whether the risk predictions from the HERO-2 based strategies are as accurate for Western patients as those from TIMI and GUSTO-I, which were derived predominantly from Western populations. We determine, using the TIMI Risk Score, whether a Western based risk strategy can be

calibrated for application to Russia, Eastern European, Latin American and some Asian countries. External validation will be performed using GUSTO-I [37] trial data. GUSTO-1 was a trial of 4 thrombolytic strategies for AMI measuring 30-day mortality in 41 021 mainly Western subjects.

We apply simple calibration methods as in Steyerberg et al [57] who previously reported encouraging results in a study comparing the performance of simple re-calibration methods to more extensive model revision.

Selected statistical measures for assessing predictive performance in validation data sets and for comparing risk indices are presented. Additionally we propose two new methods to compare the performance of competing risk models. These include a method for comparing the performance of indexes in ranking dead-alive pairs and a calibration *comparison* measure ("Entropy t-test"). The consistency of results from the various methods is assessed.

3.2 Methods

3.2.1 HERO-2 trial

The HERO-2 trial was an open-label, prospective, randomised, multicentre trial. The design and data collection methods have been described in detail previously [39]. A brief summary is provided here including details needed to provide context for this thesis.

3.2.1.1 Study population

The HERO-2 trial enrolled 17 073 patients with acute ST-elevation MI (within 6 hours of symptom onset) between November 27, 1998, and May 1, 2001. Patients of any age were eligible if they experienced chest discomfort lasting more than 30 minutes and ST-elevation or left bundle branch block on the qualifying ECG. Exclusion criteria included: active bleeding or known haemorrhagic diathesis, previous stroke, transient ischaemia attack within 6 months, current warfarin therapy, major surgery or trauma within 6 weeks, recent non-compressible vascular puncture, blood pressure of more than 180/110 mm Hg, low-molecular-weight heparin therapy within 12 hours, an activated partial thromboplastin time of at least 50 seconds in patients who had previously received heparin, and previous treatment with streptokinase. There were 1850 (10.8%) patient deaths by 30 days.

3.2.1.2 Treatments

All patients were given aspirin and then randomised via a centralised automated telephone response system to receive either bivalirudin or unfractionated heparin. Bivalirudin was given as a bolus of 0.25 mg/kg followed by an intravenous infusion for 48 hours. Heparin was given as a bolus of 5000 units followed by an infusion for 48 hours. Streptokinase was infused over 30 to 60 minutes, immediately after the bolus infusion of bivalirudin or heparin. Intravenous beta-blocker therapy was recommended for patients without contraindications, and other medications were used at the treating physician's discretion.

Data on administration of study treatments, use of aspirin or beta-blockers, coronary revascularisation and time in hospital was collected.

3.2.1.3 Geographical regions

Table 3.1 lists the 46 countries that participated in HERO-2 grouped according to their regional classification. Countries were grouped prospectively on the basis of their geographical boundaries,

Table 3.1 Countries participating in the HERO-2 trial

Region or country*	No. patients	Region or country	No. patients
Russia	**6057**	**Western countries†**	**2563**
		New Zealand	534
Eastern Europe	**5877**	Australia	462
Georgia	1153	United Kingdom	217
Poland	1108	Turkey	209
Bulgaria	867	Germany	203
Croatia	517	Netherlands	186
Ukraine	512	Belgium	170
Romania	494	Greece	164
Hungary	399	Spain	119
Czech Republic	223	France	94
Belarus	199	Canada	50
Latvia	181	Italy	49
Slovakia	168	South Africa	36
Estonia	54	United States	35
Lithuania	2	Ireland	24
		Finland	7
Asia	**756**	Portugal	3
India	565	Austria	1
Malaysia	108		
Hong Kong	47	**Latin America**	**1820**
Singapore	15	Argentina	1033
Philippines	13	Mexico	298
Thailand	8	Venezuela	227
		Chile	144
		Columbia	43
		Dominican Republic	41
		Panama	32
		Paraguay	2

*Eastern Europe, Russia and Asia together are sometimes referred to as 'Eastern regions' or 'Eastern countries' in the text.

†Western Countries included those countries in western Europe together with South Africa, Turkey, Australia, New Zealand, Canada and United States.

similar to the groupings used in other large trials (e.g., PURSUIT [79]), but Western countries were grouped together (as in InTIME-II [52] and MAGIC [80]). South Africa and Turkey were included in Western countries, and a sensitivity analysis was undertaken with these two countries excluded.

6057 (35.5%) patients were randomised in Russia, 5877 (34.4%) in other Eastern European countries, 2563 (15%) in Western countries (i.e., western Europe, Australia, New Zealand, South Africa, Turkey, Canada and the United States), 1820 (10.7%) in Latin America, and 756 (4.4%) in Asia. This method of categorising countries into regions was used throughout. Thirty-day mortality rates varied significantly according to region: Western countries (6.7%), Latin America (10.8%), Eastern Europe (10.2%), Russia (13.2%) and Asia (11.0%); $P<0.001$.

3.2.2 RISK MODELING

3.2.2.1 Baseline clinical information

Baseline characteristics available and considered for inclusion in the HERO-2 full multivariable risk model included age, sex, diastolic and systolic BP, heart rate, weight, height, and time from symptom onset to randomisation. Data were collected on an array of medical history variables including previous MI, stroke or transient ischemic attack (cerebrovascular disease), CABG or PCI, and angina. Data were also available on Killip class at entry, location of infarction by electrocardiogram, diabetes, hypertension and smoking status. Randomised treatment and region were also considered for inclusion in the model.

3.2.2.2 Missing data

Rates of missing data for the baseline variables in HERO-2 were very low (≤0.1% for all except Killip class, 1.2% missing, and height, 2.0%). Missing data occurred in 576 (3.4%) patients for at least one of the candidate baseline predictor variables. This rate was higher in Western countries (14.9%) due to increased missing data for Killip class and height. Height was not a significant factor in risk prognosis. The 30-day mortality rate was slightly higher in the subset of patients with missing data compared to patients with complete data (12.5% vs. 10.8%; $P=0.19$).

Simple imputation rules were used as described in Methods Section 2.1.2. For medical history characteristics no history was assumed. Missing Killip class and height values were estimated on the basis of multiple logistic and linear regressions, respectively, on values of all the other predictor variables. Only 11 patients had missing 30-day mortality status, of which 9 were known to have survived the index hospital admission. All 11 patients were assumed alive at 30 days.

3.2.2.3 Risk model development

For the 17 073 patients enrolled in HERO-2, univariate and multivariable analyses based on chi-square and logistic regression modeling were performed to identify predictor variables of 30-day mortality. Generalised additive models were applied in both univariate and multivariable contexts, to assess the

shape and strength of the relationship between continuous clinical variables and 30-day mortality. Consequently heart rate was fitted with a non-linear transformation applied; values below 70 bpm were truncated at 70, since the spline relationship between death and heart rate was constant for these values and values above 100 bpm were truncated at 100 since the increasing risk relationship plateaued. A backward stepwise logistic regression model was employed to identify independent predictors of 30-day mortality [81]. All eighteen baseline variables entered the initial model and were maintained if $P<0.05$. Two-way interaction effects were assessed for variables that remained in the model. To ascertain whether regional adjustment contributes significant explanatory information after adjusting for patient case-mix, the region variable was considered for inclusion in the model after independent predictor variables, and interactions between them, had been identified. Interactions between region and the baseline characteristic variables were subsequently tested to uncover differential effects in risk factors by region. The derived model is labeled the HERO-2 full risk model (HPI-FULL). The risk model linear predictor or estimated logit is calculated by taking the linear combination of regression coefficients and values of the covariables for each patient.

$$\left(\mathrm{LP} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ...\right)$$

For the convenience of clinicians a more practical and easily applied model was desirable. To obtain a more parsimonious model, we approximated the risk predictions supplied by the full risk model. Firstly Killip class categories III-IV and location of MI groups inferior and 'Other' were amalgamated. It has been established (above) that region is a significant predictor. This implies that at least simple calibration using region-specific intercepts is warranted. Furthermore, estimation of coefficients based on a model stratified by region was preferred. Stratification by region provides parameter estimates whilst controlling for the effects of region and is achieved by including region as a covariate in the model. Backward selection was then applied, with region forced into the model. Predictors were deleted (one by one) until no less than 95% of the original prognostic information was retained. In this context, R^2 statistics were evaluated as the ratio of the global χ^2 statistic (i.e., explained deviance) for reduced models compared with the full model. The 95% level proved to be appropriate as a sizeable break in R^2 values occurred below the 95% level (i.e., the removal of any additional variables would have sacrificed substantial prognostic information). The information explained by region was accounted for when calculating R^2 statistics for reduced models: only the extra deviance explained by non-regional/patient risk factors over that explained by region was considered. The resulting model is referred to as the HERO-2 reduced model (HPI-CTS).

A risk index (HPI-CAT) was then developed by grouping (in categories) all continuous variables and substituting estimated group effects. It was developed from the HERO-2 reduced model above, in consultation with clinicians to capture meaningful clinical variables and provide an index that may be readily calculated at the bed side. To derive this prognostic index, the adjusted log-odds ratio of each risk factor was divided by the smallest significant adjusted log-odds ratio, and this was rounded to the nearest integer to provide risk points associated with each risk factor. HPI-CAT produces an integer

score defined as the sum of the risk points corresponding to any given risk factor profile and classifies level of risk of death, at 30 days post MI, for a patient.

3.2.2.4 Internal validation

To enable measuring predictive performance of the HERO-2 risk strategies on independent data and to facilitate the assessment of calibrating the TIMI Risk Score in a new population we considered use of split-sample validation and resampling approaches. As stated in Methods Section 2.1.7 for EPV ≥40 optimism in performance measures has been found to be small and apparent estimates of model performance are attractive because of their stability, so these internal validation approaches were not indicated.

3.2.2.5 Risk estimation from a risk index

To estimate 30-day risk from HPI-CAT we performed a weighted logistic regression of mortality versus HPI–CAT stratified by region, to derive a risk assessment equation. Stratification by region was achieved by including region as a covariate thereby allowing region-specific adjustment of the intercept. The derived equation predicts the log odds of death; the following formula is applied to obtain the predicted risk probability: Prob[Death | HPI-CAT] = 1/[1+exp(-LP)], where LP = α_R + β HPI-CAT is the linear predictor (LP) derived from the regression coefficients α_R and β. The parameters α_R and β linearly rescale HPI-CAT before translation to the logit risk scale. In this form of simple calibration, the risk index and region adjustments when the coefficient β is common to all regions, are additive. This procedure was repeated using the HERO-2 data to relate the TIMI risk score to outcome.

3.2.2.6 Predictive performance

Table 3.2 lists a battery of methods applied in this chapter that directly compare the performance of two risk assessment strategies. A detailed description of each method now follows.

C statistics were calculated using a Wilcoxon statistic [67] which provides the same quantity as that obtained by calculating the area under the corresponding ROC curve using the trapezoidal rule. This requires forming all possible pairs of events and non-events, and counting pairs in which a risk index correctly orders the pair to derive the probability of a correct ranking. Ties are scored 0.5.

These same event non-event pairs were used to provide a further assessment of concordance and discordance between different risk indexes. Concordant results are defined as those pairs in which the HPI-CAT and TIMI indexes agree on the ordering (whether correctly or not), while discordant results were split into TIMI errors and HPI-CAT errors depending on the observed outcomes (dead or alive). When tied scores on a specific index occurred for the dead-alive individuals of any pair, the result for that index was randomly assigned as correct or incorrect. (1)

The difference in *c* statistics obtained for the HPI-CAT and TIMI indexes was evaluated and tested for significance using the test proposed by DeLong et al [82], as implemented in Stata v9.1. (2)

Table 3.2 Methods for comparing the predictive performance of risk assessment strategies

No.	Method	Aspect of performance	Basis	Reference
1	Concordant/discordant pairs	Discrimination	Errors in rankings of dead/alive pairs	Novel
2	Difference in *c* statistics	Discrimination	AUC approximated by Wilcoxon statistic.	[67], [82]
3	Net reclassification improvement	Discrimination	Risk level classification	[83]
4	Integrated discrimination improvement	Discrimination	Average sensitivity and specificity	[83]
5	"*Entropy t-test*"	Calibration	Scores calculated from deviance residuals	[84]
6	R^2 statistics on log-likelihood scale	Overall performance	Deviance explained by fitted models including risk scores as covariates.	[85]

Recently two new measures have been proposed to assess the improvement in model performance offered by the incorporation of a new biomarker. These measures are the net reclassification improvement (NRI), which is based on reclassification tables, and the integrated discrimination improvement (IDI), which is based on integrated sensitivity and specificity [83]. The NRI requires that there exist risk categories with *a priori* meaning for decision making (e.g., 0-6, 6-20, >20 percent 10-year risk of CHD based on the Third Adult Treatment Panel risk classification [86]). Individuals are classified based on risks predicted using models with and without the new marker and the two classifications are then cross-tabulated, separately for event and non-event subjects. The amount of "reshuffling", from using one model over another, is then assessed. The IDI is a continuous analogue of the NRI, i.e., no "cut-offs" are used. These measures supplement information provided by *c* statistics.

The NRI is calculated by considering the correct movement in risk categories from using one model over another – upwards for events and downwards for non-events. The improvement in reclassification is quantified as a sum of differences in proportions: the proportion of individuals moving up minus the proportion moving down for subjects who develop events; and the proportion of individuals moving down minus the proportion moving up for subjects who do not develop events. These methods were applied to compare the HPI-CAT and TIMI risk scores, calibrated, with HPI-CAT substituted as the model containing the new biomarker. So the events component of NRI was calculated as the difference in proportion of individuals assigned a higher risk category by HPI-CAT than by TIMI and the proportion classified higher for TIMI compared to HPI-CAT. The non-events component is the difference in the proportion of individuals classified lower for HPI-CAT compared to TIMI and the proportion classified lower for TIMI compared to HPI-CAT:

$$\widehat{NRI} = \left(\hat{p}_{HPI,up,events} - \hat{p}_{TIMI,up,events} \right) + \left(\hat{p}_{HPI,down,nonevents} - \hat{p}_{TIMI,down,nonevents} \right). \quad (3)$$

A simple asymptotic test is used to determine the significance of the improvement, separately for event and non-event individuals and combining the two groups (NRI).

To calculate the NRI measure we chose categories 0-<6, 6 - <20 and ≥20 per cent 30-day risk of death. Mortality in Killip class I patients, at low-risk, is around 5% in Western populations and above 30% in Killip classes III-IV, at high-risk. So the chosen categories align with the Killip class risk stratification system.

The IDI does not require categories, and focuses on the new models ability to improve integrated (average) sensitivity (IS) without sacrificing integrated (average) specificity. Sensitivity (also called recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are identified as having the condition). Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of well people who are identified as not having the condition). The integration is over all possible cut-offs. Thus the IDI quantifies jointly the overall improvement in sensitivity and specificity. The area under the sensitivity curve can be estimated by the mean of predicted probabilities of an event for those who experience events, and the area under the 'one minus specificity' curve by the mean of predicted probabilities of an event for those who do not experience events. The differences between ISs and 'one minus specificities' for the new and old model are taken:

$$\widehat{IDI} = \left(\bar{\hat{p}}_{HPI,events} - \bar{\hat{p}}_{TIMI,events} \right) - \left(\bar{\hat{p}}_{HPI,nonevents} - \bar{\hat{p}}_{TIMI,nonevents} \right). \quad (4)$$

A simple asymptotic test of significance is also available for this measure. Individual components of the IDI assessing improvement separately for IS and integrated specificity can be tested using the approach of paired samples.

The IDI and improvement in AUC are related in the sense that both are corrected average sensitivities – the IDI is corrected by the subtracted factor assessing the undesirable increase in 'one minus specificity', and the AUC by weighting the sensitivities of the two models of interest by the corresponding derivatives of specificities.

A second (new) calibration *comparison* measure ("*entropy t-test*") of two risk models was calculated [84]. This approach calculates a paired t-test to assess significance of differences in scores of two risk estimates provided by two risk models. Scores are based on the squares of deviance residuals. Residual refers to the measured discrepancy between a predicted probability and the outcome, no further model estimation is involved. Specifically, the score for individual *i* when ascribed the risk probability p_i was -2 *log* p_i if that person experienced an event (alternatively, for a non-event -2 *log*(*1-*p_i)). These risk scores can be accorded an entropy (or surprise) interpretation; the score is 0 when the prediction ascribes probability 1 to the observed event outcome, otherwise a positive score is returned.

Smaller average scores correspond to smaller observed deviance for the risk predictor and better prediction accuracy. (5)

As our final approach to compare risk models we constructed a *grouped binary dataset*, grouping by unique combinations of HPI-CAT and TIMI risk score categories and the variable region. Logistic regression models which included combinations of the HPI-CAT, TIMI risk scores, and region as explanatory variables were then fitted. Model deviances were subsequently used to calculate R^2 statistics permitting comparison of different prognostic indexes. The log-likelihood is used to provide the proportion of the null deviance explained by the fitted model (i.e., 1 – residual deviance/null deviance). These statistics were calculated for both *individual* and *group* perspectives using equations (5.10) and (5.11) of Hosmer and Lemeshow [85]. The R^2 statistics for grouped data are larger than for individuals, as the model is judged according to its ability to correctly predict group outcomes for the categories but need not correctly identify individual outcomes.

Individual and group R^2 statistics differ only in the definition of the saturated model, affecting the denominator. When R^2 is calculated at the individual level, the 'ideal' model would predict every person perfectly; that is, the predictions would be 1 and 0 for actual events and non-events. The log-likelihood of this 'ideal' model is the log-likelihood of the fitted saturated model, denoted here by $l(s)$, whereby the number of covariate patterns J equals the number of individuals n, and so $l(s) = 0$. R^2 in this individual context is calculated as: $R^2=1-l(\hat{\beta})/l(0)$ where $l(\hat{\beta})$ and $l(0)$ denote the log-likelihoods of the fitted and null models respectively. For the case $J< n$, the R^2 based on grouped data is calculated. This R^2 is computed as: $R^2=[l(0)-l(\hat{\beta})]/[l(0)-l(s)]$. Observed rates are compared with predicted rates within risk index group categories and regions. The perspective is that of a clinician providing a survival risk estimate for a member of a homogeneous group, defined by risk characteristics. Either R^2 measure requires not only relative ordering of risk ("ranking") according to different attributes, but also agreement ("calibration") of the absolute risk estimates and the observed number of events in a homogeneous risk group.

To facilitate comparison of residual deviances between models fitted to grouped and individual level data, deviances from models fitted to individual level data were rescaled by subtracting a constant. This constant is the -2LogLikelihood value for the saturated model fitted to the grouped binary dataset. (6)

3.2.2.7 External validation

The performance of the HERO-2 risk strategies and TIMI risk score (LP, calibrated using HERO-2 data) when applied to an external dataset was assessed using the intravenous–subcutaneous heparin + streptokinase arms of the GUSTO-I trial [37]. Eighty-nine patients with missing death status were excluded, resulting in a dataset containing 20 162 patients all from Western countries, except 182 patients from Poland (classified Eastern Europe). We also applied the GUSTO-I risk model to the

HERO-2 database. Application was limited to the heparin arm since the GUSTO-I model contains a term for forms of thrombolytic therapy which are incompatible with treatment with bivalirudin.

3.3 Results

3.3.1 HERO-2 multivariable models and risk score

In multivariable analysis, many of the variables that were significant in univariate analysis remained important. An exception was smoking, classified as current, past or non smoker. Current smoking was strongly associated with age and sex, and after adjustment for these variables the relative risks of smoking categories did not differ significantly. Variables excluded in backward elimination included diastolic BP, weight, height, smoking status, history of cerebrovascular disease, randomised treatment and history of CABG or PCI. There was a strong interaction between baseline variables Killip class and age. This corroborated a finding of the GUSTO-I risk model [1]: the prognostic effect of age is reduced among patients with a higher Killip class at entry. While region is statistically significant for patient outcome, no interactions of clinical importance were found between region and other risk factors. Therefore an additive model suffices, allowing the variable coefficients to be determined independent of region, by stratified regression. The model HPI-FULL introduces history of angina and sex effects, not used in the GUSTO-I model.

Variables excluded in further backward elimination were prior MI, angina, hypertension, diabetes and time from onset to randomisation. These variables were excluded because of small contributions to R^2 -- together they contributed less than 5% of total variation explained. Age was centered at 65 years to provide Killip class main effect parameter estimates at age 65. This reduced model, HPI-CTS, is presented in Table 3.3, together with the model HPI-FULL.

Scores for the further simplified HERO-2 prognostic index, HPI-CAT, range from 0 to 13. The weights are shown in Table 3.4 accompanied by the TIMI risk score, which ranges from 0 to 14. Note the omission of risk points for region and the non-differentiation of Killip classes II-IV in TIMI. Also notice the declining effect of age in HPI-CAT with increasing Killip class, as compared to risk points for TIMI.

3.3.2 TIMI risk score and HPI-CAT risk equations

The equation used to obtain a predicted probability of death within 30 days in Western Countries from the total HPI score is Prob[Death | HPI-CAT] = 1/[1+exp(-LP)], where LP = -4.226+0.404HPI-CAT. The corresponding prediction equation for the TIMI risk score, is Prob[Death | TIMI risk score] = 1/[1+exp(-LP)] where LP = -4.341+0.439TIMI. Even though the range of attainable scores is greater for TIMI than for HPI-CAT, the derived coefficients are similar, indicating HPI differentiates more between individuals. To adjust for observed geographical variation in risks a regional coefficient is added to the LP.

Adjustments necessary for region are summarised in Table 3.5, for all risk strategies, and were calculated relative to the reference class, Western countries.

Table 3.3 Coefficients of the HERO-2 full (HPI-FULL) and reduced models (HPI-CTS)

	Parameter Estimate (SE)	
Variable	**HPI-FULL**	**HPI-CTS**
Intercept:	-3.079 (0.260)	-2.920 (0.248)
Regional adjustment:		
Western countries	-	-
Latin America	0.562 (0.120)	0.641 (0.119)
Eastern Europe	0.320 (0.098)	0.402 (0.097)
Russia	0.279 (0.100)	0.452 (0.096)
Asia	1.012 (0.155)	1.018 (0.154)
Killip class*:		
I	-	-
II	0.685 (0.065)	0.740 (0.065)
III	1.306 (0.123)	1.647 (0.097)
IV	2.019 (0.158)	1.647 (0.097)
Slope of age (yrs):		
Killip class: I	0.077 (0.004)	0.080 (0.004)
II	0.050 (0.005)	0.052 (0.005)
III	0.019 (0.010)	0.013 (0.008)
IV	0.010 (0.013)	0.013 (0.008)
Systolic BP (mm Hg)	-0.018 (0.001)	-0.019 (0.001)
Heart rate (bpm)†	0.030 (0.002)	0.032 (0.002)
MI location: Anterior	-	-
Inferior	-0.509 (0.058)	-0.465 (0.055)
Other‡	-0.306 (0.118)	-0.465 (0.055)
Female	0.411 (0.060)	0.445 (0.058)
Diabetes	0.270 (0.070)	-
Hypertension	0.130 (0.061)	-
Prior MI	0.249 (0.070)	-
Prior angina	0.229 (0.062)	-
Time from onset to rx (hrs)	0.038 (0.022)	-

Rx denotes randomisation

*Killip class effects at age 65

†For heart rate, values <70 are truncated at 70 and values >100 are truncated at 100

‡"Other" MI location refers to posterior, lateral, apical, left bundle branch block or right bundle branch block

Table 3.4 HERO-2 (HPI-CAT) and TIMI risk scores [3]

Variable	HPI weight	TIMI weight
Age: <65	*	0
65-74	*	2
75+	*	3
Killip class: I	*	0
II	*	2
III-IV	*	2
Systolic BP <100 mm Hg	3	3
Heart rate >100 bpm	1	2
Anterior STE	1	1†
Female	1	–
Diabetes or hypertension or angina	–	1
Weight <67 kg	–	1
Time to treatment >4 hours	–	1

*Risk points depend non-additively on Age and Killip class, apply points in Table below.

†Anterior STE or left bundle branch block

Risk points for Killip class by Age

	HPI			TIMI		
Killip class:	I	II	III-IV	I	II	III-IV
Age:						
<65	-	3	6	-	2	2
65-74	3	5	7	2	4	4
75+	5	6	7	3	5	5

Table 3.5 Region adjustments

	Parameter Estimate (SE)			
Risk Model:	**HPI - Full**	**HPI - CTS**	**HPI - CAT**	**TIMI**
Western countries*	0	0	0	0
Latin America	0.562 (0.120)	0.641 (0.119)	0.672 (0.116)	0.473 (0.116)
Eastern Europe	0.320 (0.098)	0.402 (0.097)	0.427 (0.095)	0.348 (0.095)
Russia	0.279 (0.100)	0.452 (0.096)	0.491 (0.093)	0.388 (0.093)
Asia	1.012 (0.155)	1.018 (0.154)	1.068 (0.149)	0.558 (0.148)

*Reference

3.3.3 Comparison of the predictive performance of risk strategies applied in HERO-2

The *c* statistic of the full multivariable model was 0.82, reflecting excellent discriminatory ability. Predictive performance results for various performance measures are shown by region in Table 3.6. Model performance was marginally lower in Asia.

Table 3.6 Apparent performance of Risk Models*

Performance statistic	Overall	Western Countries	Latin America	Eastern Europe	Russia	Asia
c statistic†						
HPI - FULL	0.818	0.813	0.828	0.803	0.825	0.766
HPI - CTS	0.813	0.807	0.823	0.798	0.821	0.759
HPI - CAT	0.792	0.800	0.804	0.772	0.795	0.743
TIMI	0.781	0.787	0.794	0.761	0.780	0.761
R^2 (Nagelkerke's), %						
HPI - FULL	26.9	21.2	29.5	23.5	29.8	18.8
HPI - CTS	26.0	20.2	28.6	22.5	29.1	18.1
HPI - CAT	22.4	19.3	24.9	18.7	24.6	16.3
TIMI	20.1	17.0	23.2	16.5	21.6	18.2
Brier score						
HPI - FULL	0.080	0.056	0.076	0.078	0.091	0.086
HPI - CTS	0.080	0.056	0.077	0.079	0.092	0.086
HPI - CAT	0.083	0.056	0.080	0.082	0.095	0.087
TIMI	0.085	0.057	0.081	0.083	0.098	0.086
Calibration slope (SE)						
HPI - FULL	1	1.06 (0.081)	1.10 (0.077)	0.98 (0.042)	0.99 (0.036)	0.91 (0.115)
HPI - CTS	1	1.04 (0.081)	1.11 (0.078)	0.96 (0.042)	1.01 (0.037)	0.89 (0.113)
HPI - CAT	1	1.14 (0.090)	1.11 (0.080)	0.94 (0.043)	0.99 (0.038)	1.01 (0.130)
TIMI	1	1.07 (0.086)	1.13 (0.083)	0.93 (0.044)	1.00 (0.040)	1.11 (0.138)

*Apparent performance applying risk models to regions, within the data on which the coefficients were obtained.

†Standard errors less than 0.01 overall (in Eastern Europe and Russia ≤0.01, in Western countries and Latin America <0.02, in Asia 0.03).

The loss in predictive performance due to simplification of the full multivariable model was minor. Predictions of individual risk provided by the full and reduced multivariable models were compared for individuals (Table 3.7). Results indicate superior predictive accuracy for the full model in Eastern Europe and Russia, but failure to achieve statistical significance in other regions. However differences

Table 3.7 Entropy t-test results

Model Comparison:	HPI - FULL vs. HPI - CTS		HPI - CTS vs. HPI - CAT		HPI - CAT vs. TIMI	
	Mean	***P***	**Mean**	***P***	**Mean**	***P***
Overall	-0.005	<0.001	-0.020	<0.001	-0.013	<0.001
Western countries	-0.004	0.065	-0.004	0.460	-0.010	0.028
Latin America	-0.005	0.116	-0.021	0.005	-0.010	0.115
Eastern Europe	-0.005	0.005	-0.020	<0.001	-0.012	0.003
Russia	-0.005	0.013	-0.028	<0.001	-0.019	<0.001
Asia	-0.004	0.413	-0.009	0.532	0.010	0.365

in statistical significance appear to be attributable to sample size as the mean score differences were very similar across all regions.

The performance of the further simplified prognostic index, HPI-CAT, was marginally lower compared to the multivariable models but still demonstrated good calibration accuracy and discriminatory ability. Entropy t-tests indicate that individual risk predictions of the reduced multivariable model were superior in Latin America, Eastern Europe and Russia compared to predictions of HPI-CAT.

HPI-CAT performed better than the TIMI risk score in all regions with the exception of Asia. HPI-CAT individual risk predictions were superior to those of TIMI overall (P<0.001). The differences were not significant in Latin America and favored TIMI in Asia, although not significantly.

Estimates of calibration slope did not differ significantly from 1 for any risk strategy in any region, providing evidence that the relationship between assessed risk and outcome was common to all regions.

Figure 3.1 shows the calibration of model predictions according to deciles of predicted risk for HPI-CAT and TIMI. HPI-CAT performed better for lower risk patients.

Concordance and discordance rates between HPI-CAT and TIMI orderings of dead-alive pairs are shown in Table 3.8. Overall, error rates favor HPI, 4.3% in error, over TIMI, 5.4%. There is a high level of agreement between the two indexes, but in Western countries and Russia, particularly, HPI has a smaller error rate than TIMI. The TIMI risk score performed better in Asia.

Results for NRI are shown in Table 3.9. The Net Gain in Reclassification Proportion (NGRP) for subjects who experienced an event was not significant; 190 individuals were classified up when using HPI-CAT instead of TIMI and 184 were classified down (P=0.756). For non-event subjects, classification improved using HPI-CAT for 1860 patients, and for 850 it became worse, giving a NGRP of 0.066, significantly greater than zero (P<0.001). The NRI was estimated at 0.07 meaning that classification is improved for a net of 7% of individuals when using HPI-CAT in place of TIMI,

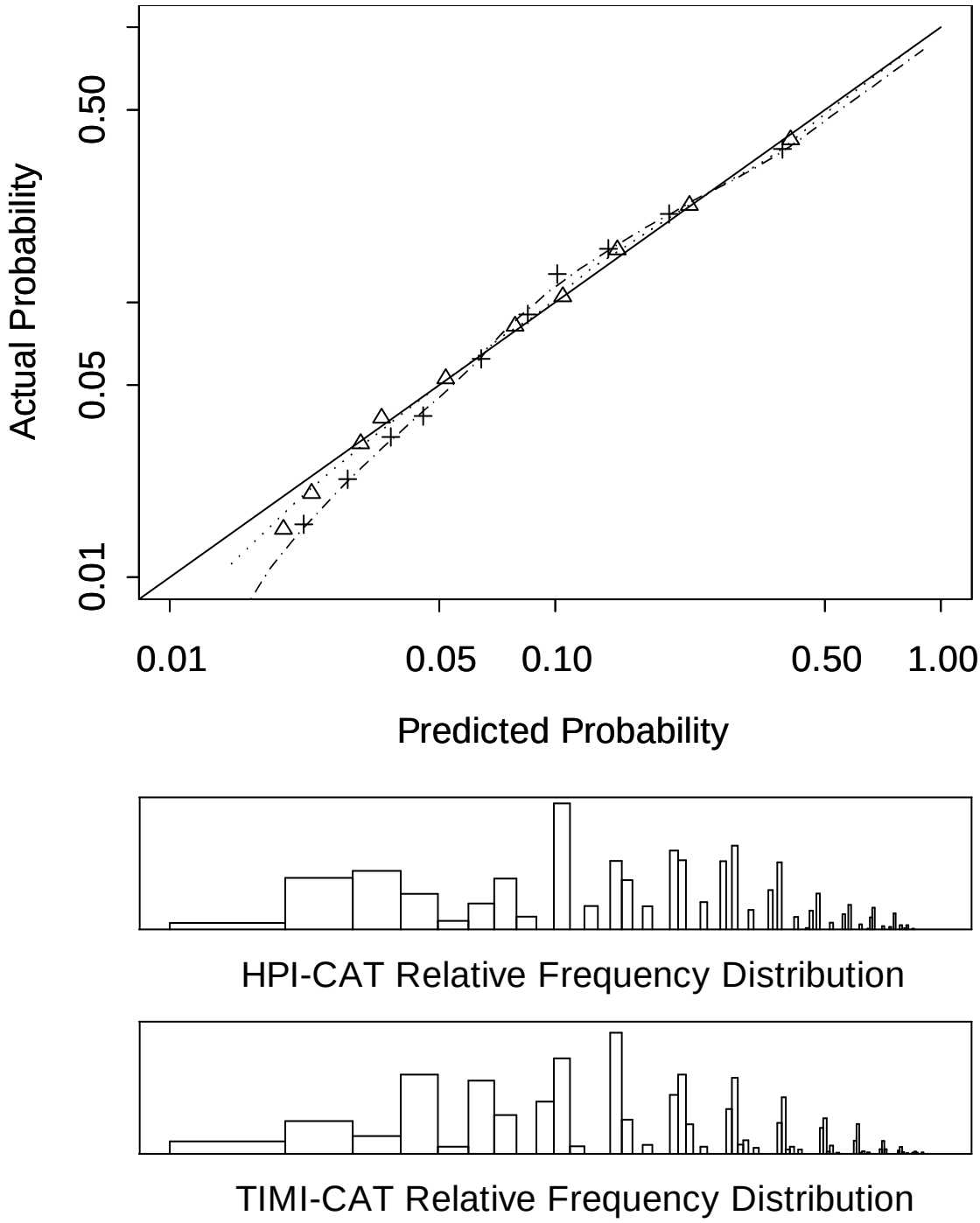

Figure 3.1 Observed vs. predicted mortality rates for the HPI-CAT (triangles) and TIMI (crosses) indexes when applied to HERO-2. The solid diagonal line reflects perfect calibration of the model predictions. Fitted smooth nonparametric lowess calibration curves are also shown (HPI = dotted line, TIMI = dashed line).

Table 3.8 Rates of agreement and disagreement of TIMI and HPI-CAT orderings of dead-alive pairs, in HERO-2

Region	Rate, %			
	Discordant			Concordant
	HPI Error	TIMI Error	Diff	
Overall	4.31	5.44	-1.13	90.3
Western countries	4.15	5.45	-1.30	90.4
Latin America	3.98	4.96	-0.98	91.1
Eastern Europe	4.67	5.78	-1.11	89.6
Russia	4.04	5.52	-1.48	90.4
Asia	7.60	5.75	1.85	86.7

Table 3.9 Net reclassification improvement results for HPI-CAT vs. TIMI (calibrated)

	Events		Non-events		Overall	
Region	**NGRP**	***P***	**NGRP**	***P***	**NRI**	***P***
Overall	0.003	0.756	0.066	<0.001	0.070	<0.001
Western countries	-0.029	0.353	0.021	0.006	-0.008	0.797
Latin America	0.015	0.590	0.060	<0.001	0.075	0.013
Eastern Europe	0.028	0.155	0.050	<0.001	0.078	<0.001
Russia	0.000	0.999	0.090	<0.001	0.090	<0.001
Asia	-0.108	0.072	0.184	<0.001	0.076	0.234

Note: Positive values favor HPI-CAT, negatives values favor TIMI.

which is highly significant. The magnitude of NRI was consistent across all regions except in Western countries, where there was no difference. There was a significant advantage with using HPI-CAT in non-event patients across all regions. In Asia the magnitude of this improvement was much larger however this was negated by a drop in classification accuracy among event patients; TIMI classified event patients marginally better than HPI-CAT. In the other regions there was no difference in performance in event patients.

Table 3.10 shows the results for IDI. Overall there was a statistically significant improvement offered by using HPI-CAT instead of TIMI. IDI was estimated at 0.017, mainly due to a 7% increase in IS (0.239 vs. 0.224 for HPI-CAT and TIMI, respectively; $P<0.001$). Within regions HPI-CAT performed significantly better in Eastern Europe and Russia. Again this improvement was mainly due to superior sensitivity for HPI-CAT over TIMI. In Asia the direction of the IDI estimate favored TIMI however the level of evidence for superiority of TIMI over HPI-CAT is weak ($P=0.810$).

Table 3.11 shows estimates of the difference in *c* statistics obtained for the HPI-CAT and TIMI indices and corresponding tests for significance. For the overall population this difference was highly

Table 3.10 Integrated Discrimination improvement results for HPI-CAT versus TIMI (calibrated)

	Events		Non-events		Overall	
Region	**IS_{HPI}-IS_{TIMI}**	***P***	**IP_{TIMI}-IP_{HPI}**	***P***	**IDI**	***P***
Overall	0.015	<0.001	0.002	<0.001	0.017	<0.001
Western countries	0.006	0.176	0.0005	0.512	0.007	0.152
Latin America	0.011	0.107	0.001	0.253	0.012	0.075
Eastern Europe	0.013	<0.001	0.001	0.024	0.014	<0.001
Russia	0.021	<0.001	0.003	<0.001	0.024	<0.001
Asia	-0.002	0.824	-0.0002	0.921	-0.002	0.810

Note: IS denotes integral of sensitivity over all possible cut-offs; IP integral of "one minus specificity".
IDI = (IS_{HPI}-IS_{TIMI}) - (IP_{HPI}-IP_{TIMI}). Positive values favor HPI-CAT, negatives values favor TIMI.

Table 3.11 Comparison of *c* statistics for HPI-CAT versus TIMI (calibrated)

	***c* statistic**			
Region	**HPI**	**TIMI**	**Difference**	***P***
Overall	0.792	0.781	0.011	<0.001
Western countries	0.800	0.787	0.013	0.173
Latin America	0.804	0.794	0.010	0.266
Eastern Europe	0.772	0.761	0.011	0.043
Russia	0.795	0.780	0.015	<0.001
Asia	0.743	0.761	-0.018	0.322

significant favoring HPI. In all regions except Asia *c* statistics increased with the use of HPI-CAT by a magnitude ranging from 0.01 in Latin America to 0.015 in Russia. This increase was significant in Eastern Europe and Russia. In Asia a decrease in discriminatory ability resulted with the use of HPI-CAT in place of TIMI; the magnitude of the difference in performance was largest for this comparison (i.e., 0.018), however was not significant.

The grouped binary dataset, generated by cross-classifying HPI-CAT, TIMI risk score categories and the variable region, produced a dataset with 358 unique observations. The results of logistic regression models, which included combinations of the HPI-CAT, TIMI risk scores, and region as explanatory variables, are summarised in Table 3.12. It is evident that overdispersion (lack of fit) is significantly less in models including HPI-CAT, compared with models with TIMI instead. As anticipated, including a region effect significantly improves the fit of both risk indexes, indicating that regional adjustment is necessary. The inclusion of an interaction term, allowing the risk score coefficient to vary by region, did not improve fit. This indicates that the strength of the association of risk with risk score does not vary according to region. TIMI, used in conjunction with HPI-CAT, adds significantly

to prediction, but only provides a small adjustment. The improvement offered by the addition of HPI-CAT to TIMI is much larger. Furthermore, models were fitted to individual level data including the LP from the multivariable model, HPI-CTS, (but excluding regional adjustments). The residual deviances of these models, which include a predictor derived from continuous and categorical variables, were substantially lower than models which included risk scores generated only from categorical variables.

Table 3.12 Results of logistic regression models, which include various combinations of the HPI-CAT, TIMI risk scores and the HERO-2 reduced model LP as explanatory variables, plus region, fitted to the HERO-2 dataset

Model	DF	Residual Deviance	*P* for dispersion	Test of nested models: *P* [model comparison]*	Group R^2, %	Individual R^2, %
1. TIMI	17071	665.19	<0.001		72.7	15.1
2. TIMI+Region	17067	639.90	<0.001	0.008 [2 vs. 1]	73.7	15.3
3. TIMI+Region+ TIMI×Region	17063	633.88	<0.001	0.508 [3 vs. 2]	74.0	15.4
4. HPI-CAT	17071	486.62	<0.001		80.0	16.7
5. HPI-CAT+Region	17067	423.20	0.005	<0.001 [5 vs. 4]	82.6	17.2
6. HPI-CAT+Region+ HPI-CAT×Region	17063	417.06	0.006	0.275 [6 vs. 5]	82.9	17.2
7. HPI-CAT+TIMI	17070	460.67	<0.001	<0.001 [7 vs. 1, 4]†	81.1	16.9
8. HPI-CAT+Region+TIMI	17066	407.78	0.020	<0.001 [8 vs. 7]	83.3	17.3
9. HPI-CAT+Region+TIMI+ HPI-CAT×Region	17062	401.96	0.022	0.285 [9 vs. 8]	83.5	17.4
10. HPI-CTS+Region	17067	76.69	NA		NA	20.2
11. HPI-CTS+Region+ HPI-CTS×Region	17063	72.50	NA	0.381 [11 vs. 10]	NA	20.2
12. HPI-CTS+Region+TIMI	17066	60.34	NA	<0.001 [12 vs. 10]	NA	20.3
13. HPI-CTS+Region+TIMI +HPI-CTS×Region	17062	56.58	NA	0.439 [13 vs.12]	NA	20.3

Notes: Models 1-9 are based on grouped binary data (see methods); models 10-13 are based on individual level data. HPI-CTS LP was fitted excluding regional adjustments.

*P-values for model comparisons were calculated using methods in section 6.5 of Collet [81].

†P<0.001 for tests comparing models 7 and 1, and 7 vs. 4.

3.3.4 Summary of HPI-CAT and TIMI risk score performance in HERO-2

The results of the comparisons of predictive performance were very consistent with both discrimination and calibration performance measures indicating modest gains in performance when using HPI-CAT over TIMI for the overall sample. Where tests of significance were available these gains were always highly significant (*P*<0.001).

The "entropy t-test", NRI, IDI and improvement in *c* statistic all indicated significant improvement in performance when using HPI-CAT over TIMI in Russia and Eastern Europe. The gains in model performance noted with the use of HPI-CAT were largest in Russia for all measures, meaning that the advantage of HPI-CAT over TIMI is at its most when applied in this region.

All methods (albeit only in "event" patients for NRI) trended towards superiority of TIMI in Asia, however results were inconclusive as the level of evidence was poor.

3.3.5 External validation in GUSTO-I trial

Table 3.13 shows the results of applying the risk strategies to the GUSTO-I data. The predicted mortality rate when the HERO-2 full model was applied was 7.2%, close to the observed rate of 7.3%. The *c* statistic of 0.81 is close to that within HERO-2.

HPI-CTS performed as well as HPI-FULL in terms of discriminatory ability but underestimated the mortality rate (6.9% vs.7.3%). Statistical assessment of individual risk predictions indicated HPI-FULL was significantly more accurate (Entropy t-test: mean score difference = -0.004; t = -4.36, $P<0.001$).

Figure 3.2 shows plots demonstrating the calibration of the model predictions according to deciles of predicted risk for HPI-FULL and HPI-CTS. The points fall very close to the line of identity between observed and expected rates for both models, slightly more for the full model, demonstrating excellent calibration.

The TIMI risk score performed better than HPI-CAT in estimating the overall mortality rate and was as accurate as the multivariable model HPI-CTS. However tests of calibration-in-the-large indicated risk predictions were on average too low for all of these approaches. The discriminatory ability of HPI-CAT was higher than TIMI based on the *c* statistic difference ($P=0.002$) and also explained a slightly higher proportion of variation (see Table 3.13). Estimates of calibration slope did not differ significantly from 1 for all risk strategies indicating neither shrinkage nor inflation of risk predictions is necessary.

The rate of concordance of HPI-CAT and TIMI orderings of dead-alive pairs was 90.6%. Among the discordant pairs the comparison of error rates favored HPI, 4.2% errors, over TIMI, 5.2%.

In terms of predicting individual risk a significant loss in predictive accuracy occurred when using HPI-CAT as an alternative to HPI-CTS ("entropy t-test": mean score difference = -0.01; t = -5.13, $P<0.001$). HPI-CAT performed significantly better than TIMI (mean score difference = -0.007; t = -3.80, $P<0.001$).

The NRI and IDI indexes were also applied to compare the HPI-CAT and TIMI risk scores and to assess consistency with other performance measures. Among subjects not experiencing an event the net gain in reclassification proportion was 0.026 favoring HPI-CAT ($P<0.001$). However for subjects experiencing an event the net gain in reclassification proportion favored TIMI and was estimated at -

0.050 (P<0.001). The NRI was estimated at -0.024 and was not significantly different from zero (P=0.06).

Table 3.13 Predictive performance of risk models applied to GUSTO-1

Performance statistic	Overall
Observed mortality, %	7.3
Predicted mortality, %	
HPI - FULL	7.2
HPI - CTS	6.9
HPI - CAT	6.7
TIMI	6.9
c statistic*	
HPI - FULL	0.809
HPI - CTS	0.804
HPI - CAT	0.787
TIMI	0.777
R^2 (Nagelkerke's), %	
HPI - FULL	22.4
HPI - CTS	21.5
HPI - CAT	19.1
TIMI	17.6
Brier score	
HPI - FULL	0.059
HPI - CTS	0.059
HPI - CAT	0.060
TIMI	0.061
Calibration slope† (SE)	
HPI - FULL	1.03 (0.027)
HPI - CTS	1.02 (0.026)
HPI - CAT	1.01 (0.027)
TIMI	1.02 (0.027)
Calibration-in-the-large: Intercept‡ (*P*)	
HPI - FULL	0.012 (0.681)
HPI - CTS	0.071 (0.014)
HPI - CAT	0.101 (<0.001)
TIMI	0.073 (0.011)

*Standard error is 0.006 for each *c* statistic.

†Both intercept and slope included as free parameters.

‡Intercept given that the slope of the linear predictor is fixed at unity.

An intercept >0 reflects that predictions were on average too low.

HPI-CAT performed significantly better for both components of IDI. The average sensitivity was 0.163 for HPI-CAT vs. 0.159 for TIMI (P=0.03) and 'one minus specificity' (i.e., the complement of specificity) was 0.060 vs. 0.062 (P<0.001). The IDI was estimated at 0.006 and was statistically significant (P=0.002); this can be interpreted as the increase in the proportion of event subjects predicted as 30-day deaths, from using HPI-CAT over TIMI, given no changes in specificity.

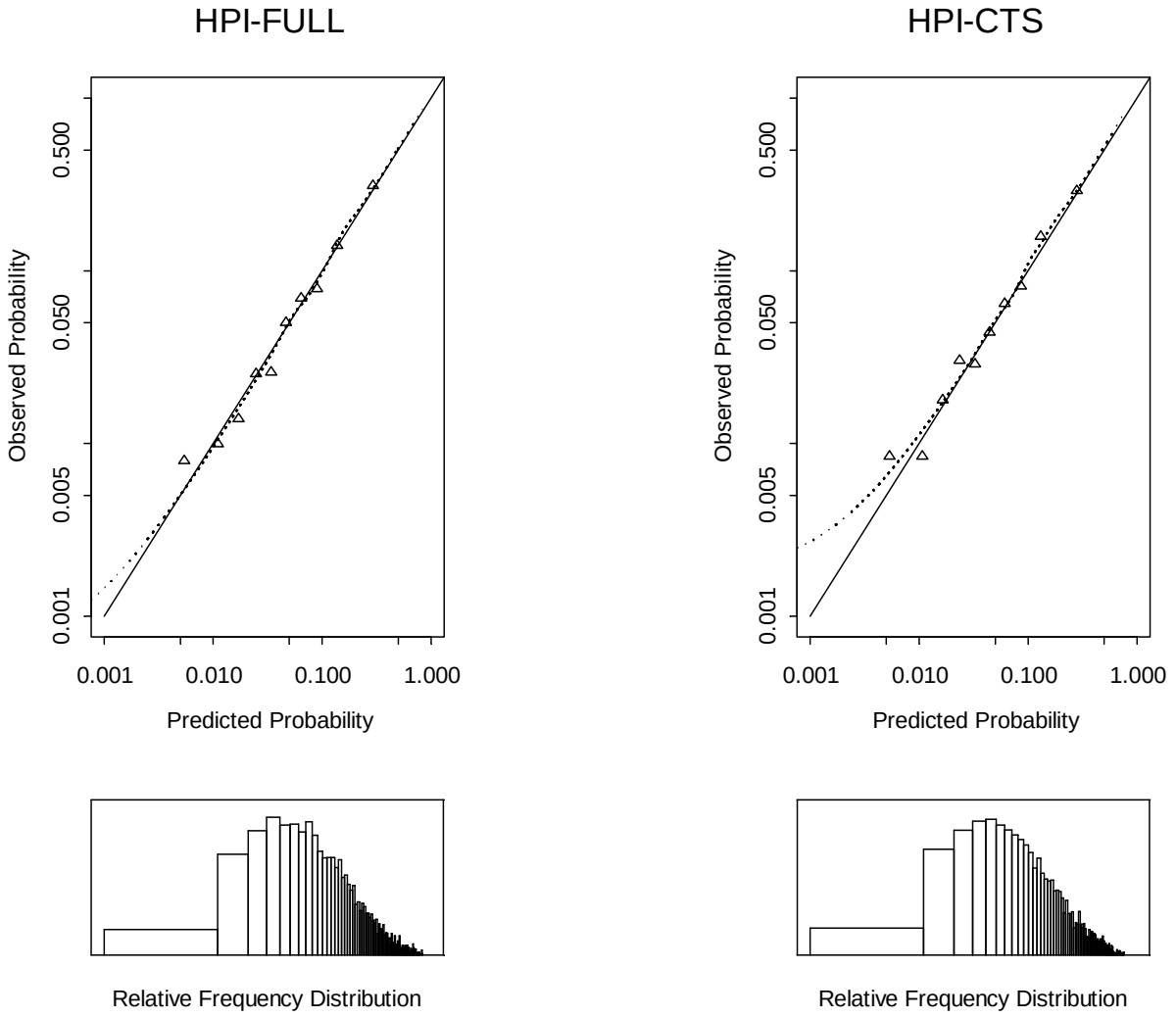

Figure 3.2 Graphs showing observed mortality vs. predicted mortality probability according to the HERO-2 multivariable full and reduced risk models when applied to the GUSTO-I data. The solid diagonal line reflects perfect calibration of the model predictions. The dotted line represents the fitted smooth nonparametric lowess calibration curve.

Figure 3.3 shows the calibration of model predictions for HPI-CAT and TIMI. In general there was little discernible difference.

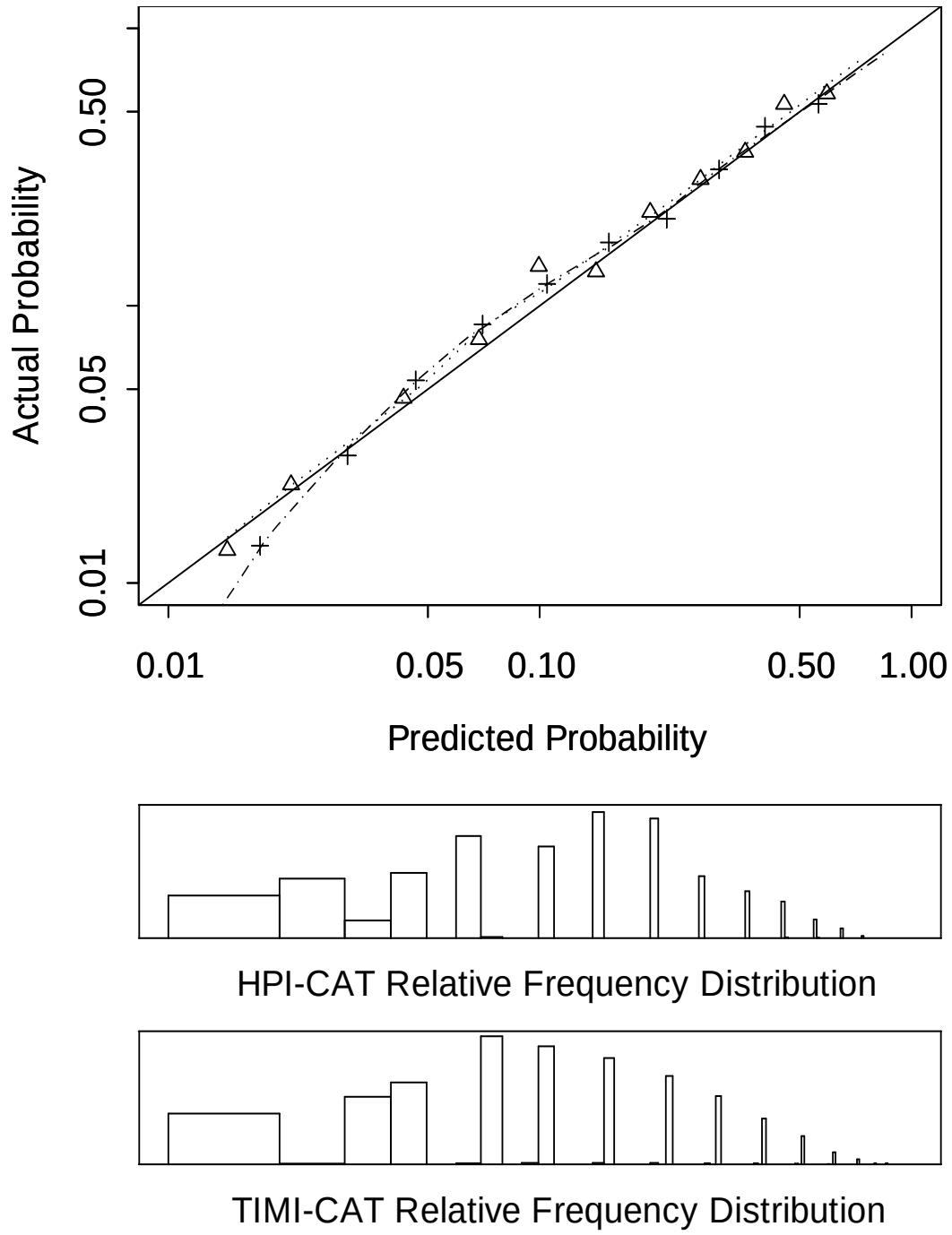

Figure 3.3 Observed vs. predicted mortality rates for the HPI-CAT (triangles) and TIMI (crosses) indexes when applied to the GUSTO-I data. Fitted smooth nonparametric lowess calibration curves are also shown (HPI = dotted line, TIMI = dashed line).

The GUSTO-I model, when applied to the HERO-2 dataset, underestimated mortality rates for the non-Western regions (Table 3.14). With the exception of Asia (c = 0.718), c statistics were similar in all regions and similar to those obtained for the full HERO-2 model.

Table 3.14 Predictive performance of the GUSTO-I model applied to the HERO-2 heparin treatment arm (n=8557)

	Mortality rate, %		
Region	**Observed**	**Predicted**	***c* statistic**
Overall	10.9	8.5	0.806
Western countries	6.7	6.5	0.828
Latin America	10.3	7.3	0.822
Eastern Europe	9.6	7.7	0.789
Russia	13.7	10.7	0.813
Asia	13.7	5.8	0.718

3.4 Discussion

We have developed a number of risk strategies for predicting 30-day mortality following hospital admission for MI. These risk assessment tools vary according to their precision of risk indication and ease of application. Choice of which index to use will depend on the setting, purpose and available resources. The risk strategies are consistent with previous risk assessment approaches including GUSTO-I [1-2], TIMI [3] and GISSI [4]. However regional differences were identified. The HERO-2 multivariable prediction model HPI-FULL, was superior at ranking risks among Asian patients compared to the GUSTO-I model [1]. This and other findings suggest that the generalisability of the HERO-2 model may be superior to that of its competitors. Also the HERO-2 model has the advantage that it is less complex and requires fewer inputs than the GUSTO-I model. The HERO-2 model itself performed very well in all aspects of predictive capability when applied to an external population (i.e., GUSTO-I).

Simplifying this multivariable model to produce HPI-CTS led to very little loss in predictive power in internal validation. Despite the higher rate of missing data in the Western region the HERO-2 models were successful in predicting risk in the Western GUSTO-I dataset. Excellent discriminatory ability was displayed, although the reduced model somewhat under-estimated the GUSTO-I observed mortality rate. Closer examination suggested the removal of the variables prior MI, angina and diabetes contributed to this underestimation. Omitting factors under-represented in the study population can lead to poorer ability to generalise to other populations. In HERO-2 the rates of prior MI (10%), angina (28%) and diabetes (13%) in Western countries were lower than in GUSTO-I (16%, 37%, and 15% respectively). As a consequence, re-calibration should be considered for risk evaluations in any new patient population, even within a single geographical region. A necessary

caveat in any use of predictive models is that extrapolation, i.e., prediction of values outside of or beyond the range of the data used to estimate the model, is not recommended.

Further simplification led to formulation of the clinical risk score HPI-CAT which can also be translated to a probability estimate. This approach was found to substantially reduce prognostic ability (R^2 statistics). The entropy t-test confirmed these findings: the largest absolute mean score differences arose from comparisons of HPI-CTS with HPI-CAT. Nevertheless, the loss in predictive accuracy may be considered acceptable as performance measures remained good. Loss in predictive accuracy might be expected as a multivariable model produces a risk probability on a continuum, from 0 to 1, whereas a risk index produces discrete values.

We have demonstrated the ability of the TIMI risk score, developed on mainly Western patients (78%), to generalise to other populations. This index performed well, when calibrated, to predict mortality rates in the HERO-2 population.

The HERO-2 study population differed from those used to develop the GUSTO and TIMI risk strategies. The 30-day mortality rate of 10.8% seen in HERO-2 is higher than the GUSTO-I [37] and InTIME-II [36] trials. Greater proportions of patients in HERO-2 were women and had anterior MI or Killip class II-IV heart failure. The time to treatment was also longer. However, the mortality rate seen in Western countries was similar. These differences were also evident when comparisons were made with other recent trials [38, 60].

Simple calibration of risk scores depends on absence of interactions of clinical importance between region and other risk factors in the population of interest, as was the case in this study. The relationship between risk assessment scores and observed mortality was consistent across geographical regions and was transportable to the GUSTO-I population. This suggests that when a risk model is (re)-calibrated for application to new populations it is only necessary to incorporate regional-specific intercepts which adjust the absolute level of expected risk. This is in agreement with the evaluation of model revision methodology of Steyerberg et al [57] which found that re-calibration of an existing index for a new environment can be as effective as more complex statistical approaches. We have deliberately not included risk points for region in HPI-CAT. Instead, simple calibration is applied in each particular region. Consequently, the risk score can be calibrated and applied elsewhere.

Both TIMI and HPI-CAT provide a large number of categories, approaching a continuum in the assessment of individual risks. This may be compared with the other extreme of a simple dichotomy of patient population into high- and low-risk, for example. However a dichotomy such as this will involve arbitrary cut-points and be difficult to calibrate in another population, where average levels of risk may change.

Risk prediction is improved by measuring age, systolic BP and heart rate precisely, rather than in categories. This was anticipated, as categorisation implies a constant effect for subjects who fall within the same category, when in fact there may be a significant gradient in response for subjects across the

category. However, discrimination or risk prediction accuracy was not substantially compromised. We felt that, when using categorisations in HERO-2, the use of clinical considerations and previous research results was a sufficient basis to determine cutpoints. By comparison, outcome-orientated methods are highly data-dependent and require further analysis to correct for multiple testing. Statistical assessment, using deciles of the relevant variable, showed the clinical cutpoints chosen to be consistent with changes in outcome rates. Potential statistical loss of power caused by categorising did not lead to important risk factors being omitted, as the risk factors which constitute the risk index were determined before this step. Despite the difficulties the approximation of the HERO-2 full model by simplified risk measures, such as cumulating risk factor "points", is useful for prognostic clinical decision making.

Statistical methods for comparisons of the discriminatory ability and calibration of two indexes have been introduced in this chapter. The tabling of concordances and discordances and the "entropy t-test" are novel and provide new information, not otherwise available. Both indicated an advantage when using HPI-CAT over TIMI. Furthermore *c* statistics were significantly higher for HPI-CAT compared to TIMI, but what constitutes a clinically important increase is unclear. Therefore we applied the recently developed measures NRI and IDI, as Pencina [83] advocates that all three quantities (i.e., improvement in *c* statistic, NRI and IDI) should yield the same conclusions to declare better performance of one model over another. The NRI and the IDI were in agreement with the *c* statistic finding in HERO-2 but, in external validation, results were less consistent, as the direction of the NRI estimate favored TIMI but was inconclusive.

The HERO-2 risk score has demonstrated that it does provide some modest improvement in predictive performance over TIMI. This improvement may be partly due to the inclusion of an age by Killip class interaction term in HPI-CAT. Further analysis revealed that the TIMI risk score over-predicted in the lowest (age<65 and Killip class I) and highest risk (age≥65 and Killip class II+) age (<65 yrs, ≥65) by Killip class (I, II/III/IV) subgroups. More detailed modeling (such as including the interaction) allows more accurate prediction than simplified modeling (such as TIMI).

To evaluate an index the recommended approach is to assess risk (on a logit scale) by a gold standard measure (here the LP of the full model) and by the pragmatic index (e.g., HPI-CAT). Model comparisons and test procedures are then available to indicate whether any statistical loss in prediction will result when the pragmatic index replaces the full model LP.

Since developing a risk model involves many steps such as defining categories, determining functional form, variable selection and pre-specifying and testing for biologically plausible interactions, validation of the resulting model in independent data is well warranted. The full procedure can be validated internally, by split-sample methods, but current research suggests this is unnecessary with large event counts [66], and so we have presented apparent estimates of predictive performance and evaluated risk models in an external, independent dataset. We conducted bootstrap resampling to

derive estimates (not included) of the expected optimism in apparent performance measures for the HERO-2 reduced model. As expected, over-optimism was negligible.

The HERO-2 risk strategies have been well validated internally, on the HERO-2 dataset, and externally, on the GUSTO-I dataset. The data necessary for external validation, however, was available only for Western populations. Since the HERO-2 strategies are based on risk factors and weightings that pertain to a global sample, we would expect high generalisability, even to non-trial settings. Calibration in each new setting using ongoing data collection is appropriate.

CHAPTER 4: INTERNATIONAL DIFFERENCES IN CLINICAL OUTCOMES AFTER ACUTE MYOCARDIAL INFARCTION

4.1 Introduction

As described in Section 1.5, worldwide mortality from acute STEMI is high [33]. This is despite treatment for this disease having improved over the last 25 years with the development of new interventions, resulting in declining fatality rates in Western countries [33, 87-93]. The HERO-2 clinical trial report revealed startling variations in 30-day mortality rates across geographical regions [39]; rates for non-Western regions were at least 50% higher compared to Western countries. Given the increasing incidence of coronary heart disease in developing countries this is concerning [33, 90, 94-96].

Using the HERO-2 trial in this chapter we explore in detail possible explanations for differences in short-term clinical outcomes between geographical regions. Since variation in mortality might be explained by differences in patient characteristics or in the use of evidence-based medicines [97-98], firstly patient characteristic data and patterns of care will be compared comprehensively across geographical regions. Secondly, randomisation rates using the HERO-2 screening log substudy will be compared; sites which randomise higher proportions of screened patients tend to have higher mortality as high-risk patients are more likely to be excluded than lower risk patients. Lastly, mortality is associated with country-level factors such as gross national income (GNI) and national health indices [49, 99]. These factors in turn are correlated with the level of development in a country and are lower in developing countries and so will also be studied by region.

The strength of the relationship between mortality and risk factors such as procedure utilisation, length of hospital stay, randomisation rates and national health and economic indicators is examined. To determine whether regional differences in mortality can be explained by these factors, the HERO-2 risk model was extended by including these factors in addition to patient characteristics. Differences in other short-term outcomes such as reinfarction, stroke and severe bleeding were also evident and hence are examined. Heterogeneity in the treatment effect for bivalirudin therapy versus unfractionated heparin according to geographical region is also studied.

Findings from the HERO-2 screening log substudy enabled assessment of the generalisability of the HERO-2 trial results and observed regional differences. Data on patient characteristics, treatments received, in-hospital mortality, randomisation status and reasons for ineligibility was recorded for 1743 patients screened for the HERO-2 trial.

4.2 Methods

4.2.1 Clinical outcomes

As already stated the primary outcome for the HERO-2 trial was mortality over 30 days. Data on other outcomes was also collected which are analyzed here. These included mortality over 24 hours, reinfarction within 96 hours, in-hospital reinfarction, stroke, and bleeding. Reinfarction within 18 hours was defined as ischemic chest pain lasting ≥30 minutes with new ST-elevation of ≥1 mm. Reinfarction occurring after 18 hours and unrelated to PCI or surgery required an elevation of creatine kinase to twice normal or a rise in the creatine kinase-MB level above the normal range, with an enzyme level at least 50% higher than the prior baseline level, or new Q waves of ≥30 ms, distinct from the enrolment MI. All reinfarctions and strokes were confirmed by clinical events committees blinded to randomised therapy.

4.2.2 Statistical analysis

Comparisons of geographic regions in baseline characteristics, treatments and outcomes were based on chi-square tests for categorical data and analysis of variance for continuous data. The Kruskal–Wallis test was used for non-normal data. The effect of regional differences on clinical outcomes used logistic regression. Covariates included in logistic models were any variables identified as significant predictors of 30-day mortality or other major outcomes. For 30-day mortality this was simply the HERO-2 full risk model. For the other clinical outcomes adjustment variables were selected by performing a backward stepwise logistic regression including the same patient-level candidate variables considered for the HERO-2 full model (as listed in Chapter 3 Methods Section 3.2.2.1). Significant predictors of outcomes were age, sex, systolic and diastolic BP, heart rate, time from symptom onset, Killip class, age by Killip class interaction, MI location, height, weight, smoking status, randomised treatment allocation and a history of MI, angina, hypertension, diabetes, cerebrovascular disease or coronary revascularisation. Variations in outcomes by region were expressed as odds ratios, with 95% confidence intervals, for each region in relation to Western countries as the reference region. The adjusted event rates were calculated for non-Western regions by application of the corresponding region group effect odds ratio obtained from the adjusted logistic regression analysis to the observed event rate.

Patterns of care and randomisation rates were included in models as country-level covariates to examine possible sources of regional variation in 30-day mortality. We calculated rates of coronary revascularisation (i.e., PCI or CABG), median durations of hospital stay and percentages of eligible patients enrolled, based on the trial data for each country. In addition, national health and economic statistics also measured at a country level were examined: GNI (estimated in US dollars per 1000 population) [100], overall health performance, and disability adjusted life expectancy (DALE) at 60 years. The latter two were based on national statistics from the World Health Report 2000 [101]. Overall health performance measures the efficiency of each country's health system and is defined as

the ratio between the actual level of health achieved in terms of gains in DALE and the maximum potential gains in DALE achievable with the available per capita health expenditure. DALE at 60 years was calculated as the average of statistics for men and women for each country, weighted in proportion to the number of men and women enrolled for that country. The strength of relations with 30-day mortality of these country-level variables was assessed unadjusted, with adjustment for patient risk factors and finally with adjustment for region and risk factors. Regional effects were examined with and without adjustment for these other variables and were considered in a multivariable model with individual risk factors and the most significant country-specific variable.

An additional analysis assessing regional effects was undertaken by using multilevel modeling to validate the findings of the conventional single-level logistic regression analysis. Hospital and country variables were also included in the model as random effects [102]. Multilevel modeling incorporates the hierarchical nature of the data which is structured into several levels: patient, hospital and country, and takes into account data that are correlated within patient subgroups. This also allows the effect of higher-level characteristics on outcomes to be assessed validly as imputing characteristics measured at higher levels to the patient-level artificially inflates the amount of available information on these variables [50]. Splus v6.1.2 for Windows was used to perform multilevel modeling.

The relative treatment effects of bivalirudin versus unfractionated heparin on mortality and on reinfarction were examined by logistic regression. Evidence of heterogeneity of treatment effect across geographic regions was examined by testing for any treatment–region interaction.

4.2.3 HERO-2 screening log substudy

We assessed a random sample of patients screened for possible enrolment in HERO-2. The aim was to ascertain the background patterns of care in the 5 regions and to compare the trial participants with those initially considered for the HERO-2 trial.

4.2.3.1 Data collection and patient population

All active sites were sent an information package inviting participation. Each site agreeing to participate was randomly assigned a period for prospective data collection during the 6 month period from September 1st 2000 to March 1st 2001. This time period was randomly assigned so that not all logs were completed during the same weeks. To provide an adequate sample for Western countries each participating site from North America, western Europe, Australia, New Zealand and South Africa were required to complete logs over a 4 week period. All other sites were only required to complete the screening log over a 2 week period. If there were major practical issues during this time period, sites were given the option of choosing a time period adjacent to the allocated time period, e.g. the 2/4 weeks immediately prior to or after.

All patients presenting during the specified period with suspected acute evolving STEMI were recorded on the log regardless of the treatment the patient subsequently received.

Information was collected on patient demographics (i.e., age and sex) and treatment received (i.e., reperfusion and anti-thrombin therapies administered). For patients randomised to HERO-2 the study number was recorded. For patients not randomised a coded reason was provided. For those patients whose reason was exclusion criteria met, the actual criteria met were also noted. Discharge mortality status was also recorded.

4.2.3.2 Sample size

Based on a recruitment rate of 150 patients per week it was envisaged that the substudy should provide approximately 300 randomised patients and 500-800 non-randomised patients worldwide. A study size of 1000 patients would enable differences in proportions of 10% to be detected with at least 80% power. The eventual sample size was 1743 screened patients of which 274 were randomised.

4.2.3.3 Analysis

Patients who were admitted outside of the correct screening dates were excluded from all analyses. For the sites which did not conduct the screening log during the allocated time period or during an adjacent period, patients who were screened after the 2/4 weeks from the date the site commenced screening were excluded. This was necessary for comparisons of screening and recruitment rates (not reported here).

The HERO-2 trial database was used to fill in missing data on the screening log for randomised patients.

4.3 Results

4.3.1 Baseline characteristics

Table 4.1 shows baseline characteristics by region. Of the variables considered the most clinically important a higher proportion of elderly patients (age ≥75 years) were recruited in Western countries however patients in this region presented earlier (less with symptom onset >4 h) and were less likely to have anterior MI.

Higher proportions of females and patients with higher Killip class (i.e., ≥II) were recruited in Eastern Europe and Russia. Russia was also associated with higher proportions of patients having a history of cardiovascular disease, specifically a prior MI or angina.

The prevalence of hypertension was higher in Latin America, Eastern Europe and Russia.

Patients from Asia were significantly smaller (by weight or height) and strikingly younger and less likely to be female than those from other regions. Accordingly lower proportions of patients had suffered a prior MI or angina.

Overall risk in Western countries, as measured by the GUSTO-I [2], and TIMI [3] risk scores was similar to, or as high as, that in other regions.

Table 4.1 Characteristics of HERO-2 patients, by region[*]

Region	Western Countries	Latin America	Eastern Europe	Russia	Asia
No.	2563	1820	5877	6057	756
Age (years)[‡]	63 (53-73)	60 (51-69)	62 (52-70)	63 (52-71)	55 (46-62)
Age ≥75 years (%)	18	10	12	13	3.8
Female (%)	25	22	29	33	14
Previous history (%)					
angina	28	38	44	63	22
myocardial infarction	10	8.9	13	23	2.7
cerebrovascular disease	2.3	1.3	0.9	1.5	0.8
coronary revascularisation	5.5	4.4	1.0	0.6	0.7
Risk factors					
current smoker (%)	43	47	43	43	51
hypertension (%)	36	51	52	62	28
diabetes (%)	13	19	16	10	22
anterior MI (%)	32	43	44	51	51
symptom onset >4 hours (%)	20	32	26	32	40
weight (kg) [‡]	76 (69-85)	76 (68-85)	78 (70-87)	75 (67-84)	60 (55-68)
height (cm) [‡]	170 (165-176)	168 (162-174)	170 (165-176)	169 (163-175)	160 (157-167)
Hemodynamics[‡]					
systolic blood pressure (mm Hg)	134 (120-150)	130 (120-150)	138 (120-150)	131 (120-150)	130 (113-150)
diastolic blood pressure (mm Hg)	80 (70-90)	80 (70-90)	80 (72-90)	80 (75-90)	80 (70-90)
heart rate (beats/min)	72 (61-84)	76 (65-88)	77 (66-90)	76 (66-88)	82 (71-92)
Killip class II (%)	12	12	16	24	13
Killip class III or IV (%)	1.4	1.9	3.3	6.7	1.2
GUSTO-I risk score[‡§]	60 (50-70)	58 (48-68)	61 (51-71)	62 (51-74)	55 (46-64)
TIMI risk score[‡¶]	3 (1-4)	3 (2-4)	3 (1-5)	4 (2-5)	3 (2-4)

[*] $P<0.0001$ for regional differences in all variables. [‡]Median (interquartile range). [§]See Califf et al [2]. [¶]See Morrow et al [3].

4.3.2 Patterns of care

The quality of therapy was high for all regions: 99% of patients received their treatment according to the protocol, and its duration was 48 hours for most (Table 4.2). Intravenous beta-blocker therapy was given to a minority and was highest in Western countries and Latin America. Other concomitant medications were not recorded.

Large disparities existed by region in the use of procedures and other care. The rates of angiography and PCI were low overall however use of these procedures was much higher in Western countries and very low in Eastern Europe and Russia.

The median duration of hospital stay and the autopsy rate was higher in Eastern Europe and Russia, particularly in Russia. In contrast, the use of CT scanning and MRI machines among patients suffering stroke was much lower in Russia.

4.3.3 Clinical outcomes by region

As stated the overall thirty-day mortality rate was 10.8%, but was significantly lower in Western countries, at 6.7%, than in the other regions (Figure 4.1). This was higher by 4.1% (95% confidence interval (CI) 2.4%–5.8%) in Latin America, 3.5% (95% CI 2.3%–4.8%) in Eastern Europe, 6.5% (95% CI 5.2%–7.8%) in Russia, and 4.3% (95% CI 1.9%–6.7%) in Asia.

Adjustment for prognostic factors reduced these differences for Eastern Europe and Russia as shown by the adjusted event rates; however the increased mortality was still significant as the odds ratio CIs were entirely above 1. For Latin America and Asia the effect of the adjustment was the reverse; for Asia the adjusted event rate increased by roughly 50%.

In summary the variation in mortality by region was unexplained by differences in known patient risk factors. After adjustment for baseline risk factors regional variation was still highly significant (P<0.001) with all non-Western regions having significantly increased mortality relative to Western countries.

A similar pattern was displayed for 24-hour mortality.

Rates of reinfarction, during the in-hospital period and within 96 hours of randomisation, varied across regions. Rates were significantly higher in Western countries than in Latin America, Eastern Europe and Russia, even after adjustment for baseline risk factors. Again the event rate for Asia increased dramatically after adjustment but was not significantly different compared to Western countries. Adjusted odds ratios for reinfarction within 96 hours and during hospital stay were similar.

The overall variation in stroke rates by region was significant, even after adjustment for baseline risk factors (P=0.02). Rates in Russia and Asia were higher, while rates in Western countries and Eastern Europe were lower.

The variation in severe bleeding rates by region was significant (P<0.001). Compared to Western countries rates were significantly lower in Eastern Europe and Russia and significantly higher in Asia.

Table 4.2 Patterns of care and protocol treatment by region in the HERO-2 trial[*]

Region	Western Countries	Latin America	Eastern Europe	Russia	Asia
No.	2563	1820	5877	6057	756
Drug administration					
Time from randomisation to thrombolytic treatment (min)[‡]	15 (11-23)	19 (13-27)	12 (10-17)	13 (10-18)	15 (10-22)
Duration of thrombolytic treatment (min)[‡]	58 (40-60)	51 (39-60)	60 (45-60)	45 (37-58)	55 (40-60)
Duration of antithrombin (hours)[‡]	48 (48-49)	48 (48-48)	48 (48-49)	48 (48-48)	48 (48-49)
Streptokinase (%)	98	99	100	100	99
Aspirin (%)	99	99	99	99	100
Intravenous beta-blocker (%)	16	16	13	11	11
Other care					
Angiography (%)	35	26	6.7	1.4	20
PCI (%)	18.3	10.3	3.4	0.3	9.9
CABG (%)	2.7	1.8	0.2	0.1	0.8
Duration of hospital stay (days)[‡]	7 (5-11)	7 (6-11)	13 (10-17)	22 (18-26)	6 (4-8)
CT or MRI (% of stroke patients)	100	85	73	34	92
Autopsy rate (% of in-hospital deaths)	6	7	46	79	3

[*]$P<0.0001$ for regional differences in all variables. [‡]Median (interquartile range). PCI=percutaneous coronary intervention also known as angioplasty; CABG=coronary artery bypass grafting; CT=computed tomography; MRI=magnetic resonance imaging.

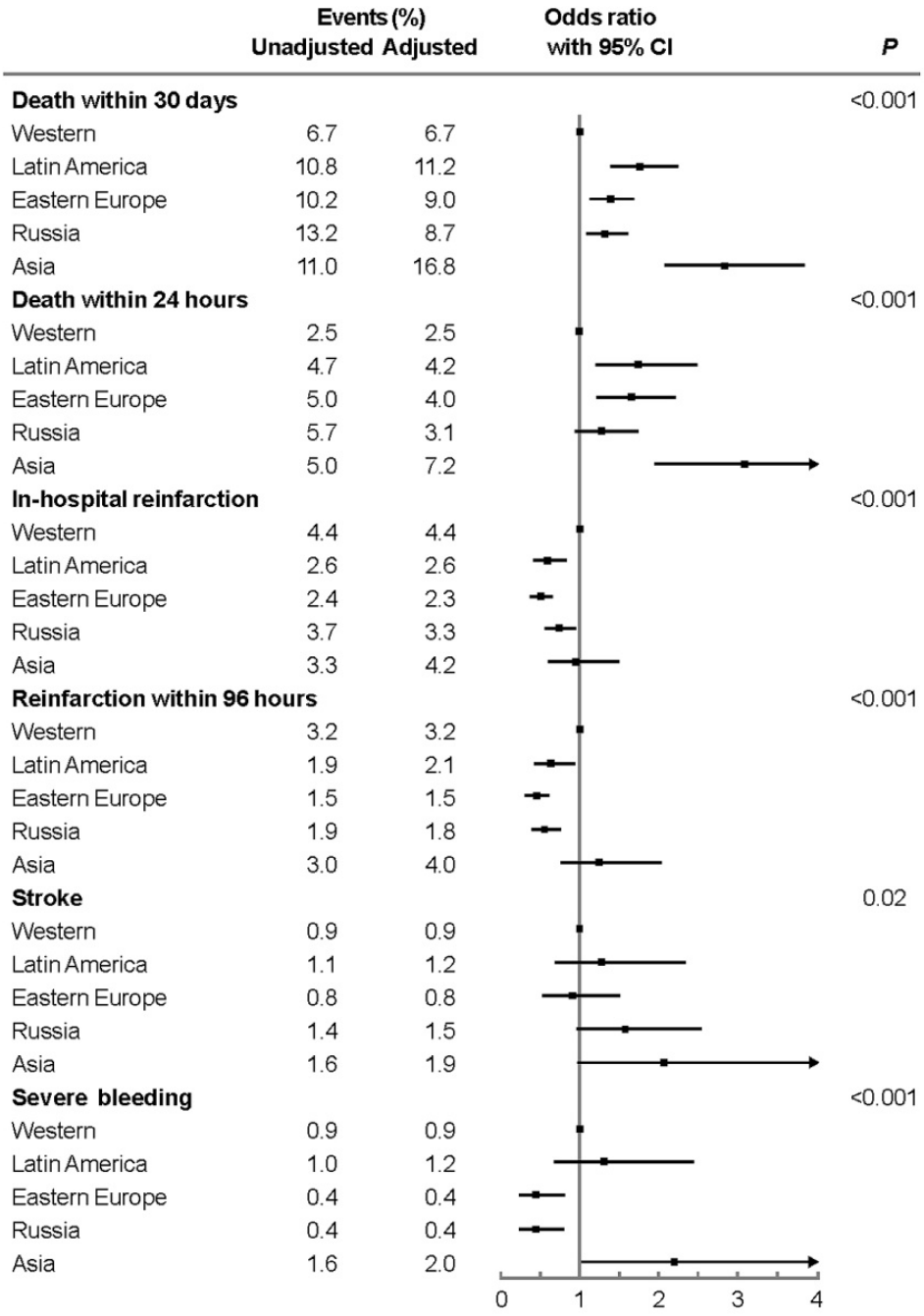

Figure 4.1 Clinical outcomes by geographic region: Western countries, Latin America, Eastern Europe, Russia and Asia. Adjusted rates are expressed as relative to the rate for Western countries, adjusted for baseline prognostic factors. *P*-values are for the regional effect from a multivariable model adjusted for those prognostic factors that were significant for each outcome (from the following: age, sex, systolic and diastolic BP, heart rate, time from symptom onset, Killip class, age by Killip class interaction, MI location, height, weight, smoking status, randomised treatment allocation, and a history of MI, angina, hypertension, diabetes, cerebrovascular disease or coronary revascularisation).

4.3.4 HERO-2 screening log substudy results

Of the 538 sites in the HERO-2 study 268 (50%) agreed to participate in the screening log substudy. Of these, logs were received from 222 sites. For 12 sites zero patients were screened. A total of 1774 patients were screened, of which 277 (16%) were randomised.

Fifteen sites screened patients after or before the correct screening period; 31 patients have been excluded as a result, 3 of which were randomised. This leaves 1743 screened patients at 222 hospitals (42% of all sites and 74% of all countries) for analysis, of which 274 were randomised. Of screened patients 699 met all inclusion and none of the exclusion criteria and hence were classified as eligible for the trial.

Table 4.3 summarises characteristics for screened patients (N=1743), eligible patients (N=699) and reasons patients were not randomised (N=1469).

For screened patients the pattern of regional variation was similar to in the whole HERO-2 trial. Eastern Europe and Russia had higher proportions of females and patients from South America and Asia were younger. Western countries had a much lower mortality rate compared to all other regions (P<0.001). Rates of utilisation of reperfusion therapy were substantially lower in Eastern Europe and Russia (less than 50%). Lastly higher proportions from Eastern Europe and Russia than from Western countries were randomised (P<0.001).

Trends by region were similar for eligible patients. The use of PCI and other lytics was more common in Western countries and Latin America, which partly explains the lower rates of randomisation in these regions. The significant regional variation in mortality was again evident (P=0.005).

Patients in Eastern regions were more often not randomised due to not meeting the eligibility criteria; these patients presented late (>6 hours) more often than patients from Western countries and Latin America.

4.3.5 Other explanatory regional variables

Table 4.4 shows the results of univariate and multivariable analysis examining the association between country-specific factors and region with 30-day mortality. Table 4.5 shows summary statistics for national health indicator variables by region.

In univariate analysis countries with higher GNI, overall health system performance and DALE at 60 were associated with lower 30-day mortality (each P<0.001), as were countries with higher rates of coronary revascularisation (Tables 4.4 and 4.5). Countries with longer hospital stays and higher rates of eligible patients randomised were associated with increased mortality. After adjustment for individual baseline risk factors only DALE at 60 remained significant; for each year increase in life expectancy the odds of death decreased by 4%. Further adjustment for region produced nonsensical results, particularly for overall health system performance. This is probably related to the strong association between overall health system performance and region.

Table 4.3 Characteristics and outcomes of a sample of patients recruited and screened in the HERO-2 trial (*N*=1743)

Region[*]	**Western Countries**	**Latin America**	**Eastern Europe**	**Russia**	**Asia**
n =	507	112	464	589	71
Age: median (lower quartile - upper quartile)	67 (55-76)	59 (51-68)	65 (56-72)	66 (55-74)	59 (50-69)
Female (%)	28	29	36	40	21
Reperfusion therapy (%)					
thrombolysis	61	47	36	26	68
acute PCI	6	14	2	0	0
neither	33	39	62	74	32
Randomised (%)	7	10	20	21	15
Died in hospital (%)	7.3	16.4	12.3	16.5	14.3
Eligible patients (N = 699, 40%); n =	262	59	165	183	30
Age: median (lower quartile - upper quartile)	66 (54-74)	58 (50-68)	63 (53-72)	62 (52-72)	60 (52-70)
Female (%)	25	20	33	32	20
Reperfusion therapy (%)					
streptokinase	42	46	69	77	90
other lytic	40	15	7	0	0
acute PCI	10	19	3	1	0
none	8	20	21	22	10
Randomised (%)	13	19	56	69	37
Died in hospital (%)	8.5	20.7	14.3	16.9	26.7
Reasons 1469 patients not randomised (%)					
No pain or ECG criteria	10	16	8	30	2
Onset >6 hours	26	27	48	38	52
Exclusion criteria	16	9	24	20	15
Eligible: decision of doctor or patient	37	31	17	8	28
Eligible: other reason[‡]	11	17	3	4	3

[*]$P<0.001$ for all regional differences except among eligible patients: sex ($P=0.1$), age ($P=0.01$) and deaths in hospital ($P=0.005$). [‡]Other reasons were logistical reasons, enrolled in other trial, or early death or cardiac arrest. PCI=percutaneous coronary intervention.

Table 4.4 Effect of risk score and country factors on 30-day mortality and regional variation in the HERO-2 trial

	Univariate analysis		Adjusted for risk factors		Adjusted for region and risk factors	
Country-specific factors[*]	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*
GNI (÷ 1000)	0.97 (0.97-0.98)	<0.001	0.99 (0.98-1.00)	0.07	1.02 (1.00-1.04)	0.05
Overall health system performance (x 100)	0.99 (0.98-0.99)	<0.001	1.00 (1.00-1.01)	0.49	1.02 (1.01-1.03)	<0.001
DALE at 60 years	0.92 (0.89-0.94)	<0.001	0.96 (0.93-1.00)	0.03	1.03 (0.96-1.12)	0.41
Revascularisation rates	0.98 (0.98-0.99)	<0.001	1.00 (0.99-1.01)	0.98	1.01 (1.00-1.01)	0.28
Duration of hospital stay	1.03 (1.02-1.04)	<0.001	1.00 (0.99-1.01)	0.64	1.03 (1.01-1.05)	0.01
Randomisation rate (% eligible)[†]	1.01 (1.00-1.01)	<0.001	1.00 (1.00-1.00)	0.10	1.00 (0.99-1.00)	0.01
	Univariate analysis		**Adjusted for risk factors**		**Adjusted for country-specific factors and risk factors**[‡]	
Region	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*	Odds ratio (95% CI)	*P*
Western Countries	1.0	<0.001	1.0	<0.001	1.0	<0.001
Latin America	1.69 (1.36-2.09)		1.76 (1.39-2.23)		2.37 (1.81-3.11)	
Eastern Europe	1.59 (1.34-1.90)		1.38 (1.14-1.68)		2.07 (1.60-2.67)	
Russia	2.13 (1.79-2.53)		1.33 (1.09-1.61)		2.80 (1.92-4.08)	
Asia	1.73 (1.31-2.27)		2.82 (2.08-3.82)		4.72 (3.24-6.88)	

[*]Odds ratios for country-specific factors are expressed per unit change of each variable. For example, 0.97 for GNI (÷ 1000) corresponds to a 3% reduction in the odds of 30-day mortality for each $1000 increase in GNI.

[†]Calculated from assessment of patients recruited and screened (Table 4.3).

[‡]Multivariable model including region, patient risk factors and single most significant country-specific factor: overall health system performance.

GNI = Gross national income per capita; DALE = disability-adjusted life expectancy at age 60 (average of values for men and women weighted by enrolment of each sex in each country).

Table 4.5 National health statistics by region. Weighted average of national health statistics and trial variables by region

Factor	Western Countries	Latin America	Eastern Europe	Russia	Asia
Gross national income (per capita, $US)*	18053	6138	2538	1690	2959
Overall health system performance (x100)†	87.4	75.1	70.5	54.4	64.7
Disability-adjusted life expectancy at birth (years) †	69.3	64.7	63.5	59.5	55.1
Disability-adjusted life expectancy at 60 (years) †	16.3	15.2	13.4	11.9	10.8

*World Bank, 2002[100]

†World Health Report, 2000[101]

A series of models was also performed which included region and each country-specific variable; 30-day mortality still varied significantly across the 5 regions (each $P<0.001$). Regional variation also remained highly significant after adjustment for all patient risk factors and any of these country-specific variables ($P<0.001$).

4.3.6 Multilevel modeling

Table 4.6 shows results for the above analysis repeated using multilevel models with hospital and country variables included as random effects. As expected the strength of associations with mortality for country-specific variables reduced as the amount of available information on these variables was no longer inflated as they were incorporated correctly into the model as variables measured at the country-level not at the patient-level as in the single-level analysis. In univariate analysis only GNI, coronary revascularisation and percent randomised remained significant with smaller effect estimates compared to in the standard logistic analysis. A much larger proportion of the variability in mortality was attributed to variation between hospitals as shown by the inter-hospital and inter-country variances.

After adjustment for patient risk factors none of these variables remained significantly associated with mortality and the allocation of residual variance attributed to unmeasured characteristics at the hospital and country levels changed. The variability in outcome attributed to hospital variation decreased by around 38% and the inter-country variances increased 2-3 fold becoming larger than the inter-hospital variances.

After adjusting for region almost all of the variation in mortality between countries was accounted for. Health system performance was significantly associated with mortality however as in the standard logistic analysis the direction of the effect was not consistent with prior expectations.

Regional variation remained highly significant in univariate analysis and after adjustment for patient risk factors and country-specific factors ($P<0.001$). Effect sizes for region group effects were similar in the standard logistic analysis although CIs were wider as multilevel models take into account the correlated nature of the data. Notwithstanding, all non-Western regions were still associated with significantly higher odds of mortality compared to Western countries as CIs did not overlap one.

4.3.7 Treatment effects by region

The 30-day mortality event rates were similar between the bivalirudin and heparin treatment arms for the overall cohort (Odds ratio (OR) = 0.96 (95% CI 0.86-1.07), $P=0.46$). Within regions there was a significant benefit with the use of bivalirudin instead of heparin in Asia ($P=0.02$), however the overall test for heterogeneous treatment effects across regions was not statistically significant (Figure 4.2).

Bivalirudin was associated with a lower rate of adjudicated reinfarction than heparin for both the 96 hour endpoint (OR = 0.70 (95% CI 0.56-0.87), $P=0.001$), and during the hospital stay (OR = 0.77

Table 4.6 Multilevel modeling results: effect of risk score and country factors on 30-day mortality and regional variation in the HERO-2 trial

	Univariate analysis				Adjusted for risk factors				Adjusted for region and risk factors			
Country-specific factors*			Variance				Variance				Variance	
	OR (95% CI)	*P*	Hospital	Country	OR (95% CI)	*P*	Hospital	Country	OR (95% CI)	*P*	Hospital	Country
GNI (÷ 1000)	0.98 (0.97-0.99)	0.01	0.115	0.021	0.99 (0.97-1.00)	0.15	0.072	0.069	1.02 (0.99-1.04)	0.06	0.079	0.0001
Overall HSP (x 100)	0.99 (0.99-1.00)	0.11	0.116	0.026	1.00 (0.99-1.01)	0.82	0.071	0.082	1.02 (1.01-1.03)	<0.001	0.063	0.00005
DALE at 60 yrs	0.96 (0.91-1.01)	0.13	0.115	0.032	0.96 (0.90-1.02)	0.17	0.071	0.066	1.10 (0.99-1.21)	0.06	0.085	0.0008
REVAS rates	0.99 (0.98-0.99)	0.04	0.115	0.032	1.00 (0.99-1.01)	0.54	0.071	0.083	1.00 (0.99-1.01)	0.36	0.077	0.0001
LOS	1.02 (0.99-1.04)	0.07	0.113	0.028	0.99 (0.97-1.02)	0.70	0.070	0.083	1.02 (0.99-1.05)	0.07	0.069	0.00005
RX rate (% eligible)†	1.00 (1.00-1.01)	0.03	0.116	0.035	1.00 (0.99-1.01)	0.65	0.070	0.086	1.00 (0.99-1.00)	0.11	0.070	0.00005

	Univariate analysis				Adjusted for risk factors				Adjusted for country-specific factors and risk factors‡			
Region			Variance				Variance				Variance	
	OR (95% CI)	*P*	Hospital	Country	OR (95% CI)	*P*	Hospital	Country	OR (95% CI)	*P*	Hospital	Country
Western Countries	1.0	<0.001	0.108	0.00005	1.0	<0.001	0.078	0.0001	1.0	<0.001	0.063	0.00005
Latin America	1.68 (1.32-2.15)				1.75 (1.36-2.26)				2.36 (1.76-3.16)			
Eastern Europe	1.56 (1.27-1.92)				1.43 (1.16-1.77)				2.08 (1.56-2.76)			
Russia	1.90 (1.53-2.36)				1.28 (1.01-1.59)				2.67 (1.73-4.12)			
Asia	1.73 (1.22-2.44)				2.85 (1.99-4.02)				4.67 (3.06-7.13)			

*Odds ratios for country-specific factors are expressed per unit change of each variable. For example, 0.98 for GNI (÷ 1000) corresponds to a 2% reduction in the odds of 30-day mortality for each $1000 increase in GNI.

†Calculated from assessment of patients recruited and screened (Table 4.3).

‡Multivariable model including region, patient risk factors and single most significant country-specific factor: overall health system performance.

GNI = Gross national income per capita; HSP = health system performance; DALE = disability-adjusted life expectancy at age 60 (average of values for men and women weighted by enrolment of each sex in each country); REVAS = revascularisation; LOS = length of hospital stay; RX = randomisation; OR = Odds Ratio.

(95% CI 0.65-0.92), *P*=0.004). The treatment effects were similar across all regions for both outcomes.

There was no evidence of heterogeneity in treatment effects across regions for the outcomes stroke (*P*=0.70) and severe bleeding (*P*=0.63).

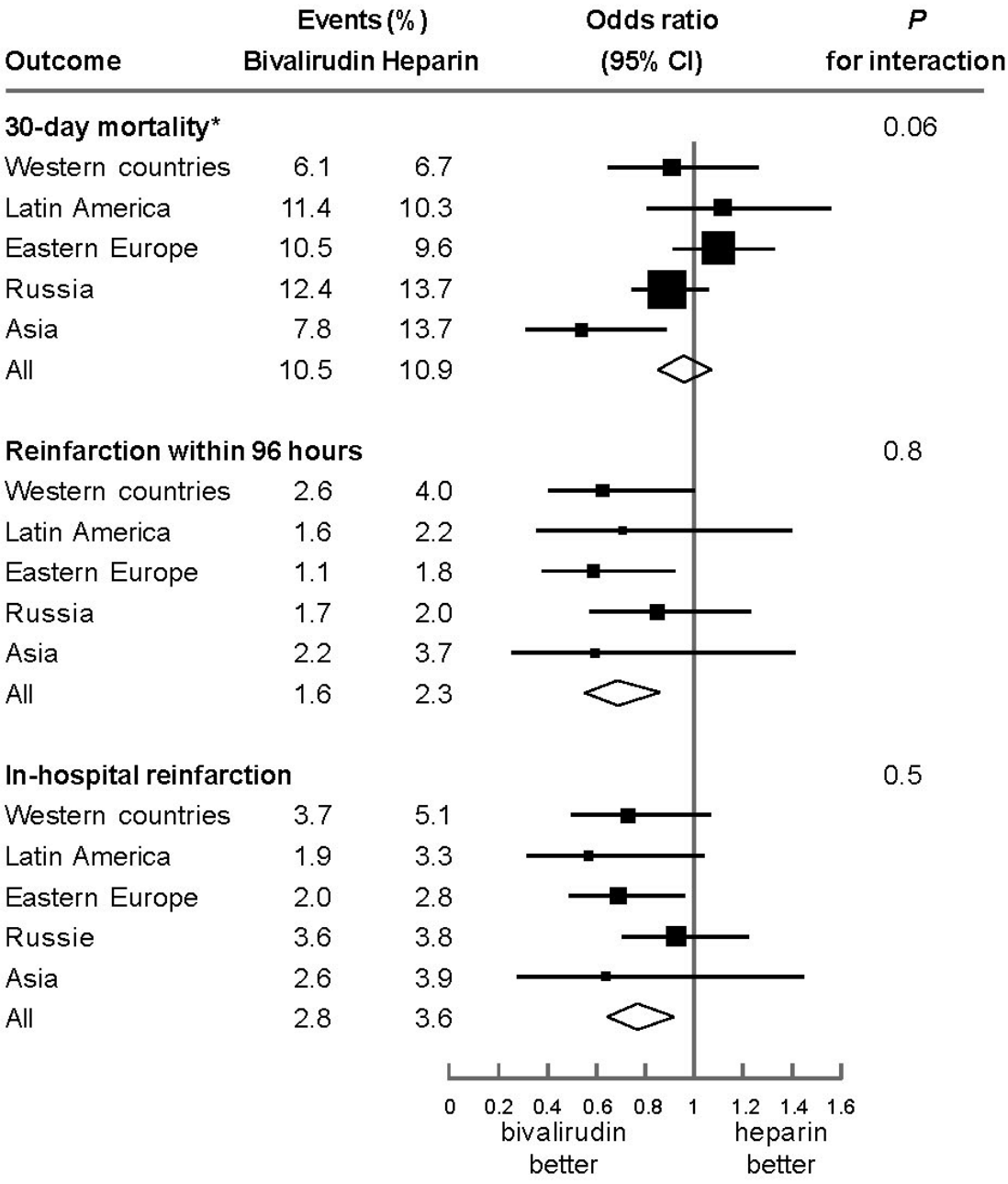

Figure 4.2 Effects of bivalirudin versus heparin according to region. Odds ratio of treatment effect, with adjustment for baseline prognostic factors.

4.4 Discussion

In the HERO-2 trial we observed large variations in 30-day mortality rates across regions. We were unable to explain the differences using patient characteristics, patterns of care and national health and economic statistics. Significant variations were also observed for other outcomes such as reinfarction, stroke and severe bleeding which were also not explained. Our findings provide strong evidence that regional variation in the outcomes of AMI is real as we recruited large numbers in non-Western regions, more than in all previous studies. This work has not only confirmed earlier studies for regions previously studied [52, 80] but has extended them due to the added geographical diversity of the HERO-2 sample. The diversity in outcome rates is more extensive than previously realised with poorer outcomes also evident in Asia.

The 30-day mortality rates seen in HERO-2 were similar to other trials in Western countries and Latin America both in trials of streptokinase and the newer thrombolytic agents [37-38, 52, 60-62]. Only two other trials in AMI have also recruited large numbers in Eastern Europe and Russia, namely InTIME-II [52] and MAGIC [80]. For these regions mortality was around 3% higher in MAGIC than in HERO-2, however MAGIC patients were on average six years older and only 20% received reperfusion treatment (i.e., thrombolytics or PCI) compared to 100% in HERO-2. In InTIME-II mortality was lower (7.3%) however patients had lower risk-factor profiles. The GUSTO-3 trial recruited 886 patients in Eastern Europe with similar risk-factor profiles to patients in HERO-2. The corresponding mortality rate was more similar to HERO-2 although not as high (8.7 vs. 10.2%).

Similar regional trends were observed for other outcomes in InTIME-II. Bleeding rates varied significantly and were lowest in Eastern Europe and highest in Latin America. Non-fatal reinfarction rates were also highest in Western countries and lowest in Latin America.

There were differences in patient risk factor profiles across regions; similar trends were seen in the MAGIC trial [80]. Most notably Western patients were on average older but this was negated by more favorable risk factors such as earlier presentation time and non-anterior MI. Also of critical clinical relevance there were more patients with Killip class ≥II in Eastern Europe and Russia. However differences in known risk factors did not account for the regional variation as these factors were adjusted for in the multivariable analysis. There may be differences in unknown risk factors not measured in HERO-2 such as ethnicity, lifestyle or genetic factors which may have contributed to the observed variation.

There were differences in the patterns of care across geographical regions. Most relevant the use of revascularisation was substantially lower in non-Western regions and more delayed (data not shown) in Eastern Europe and Russia, consistent with other studies [52, 80]. We adjusted for differences in the use of revascularisation at the country level (i.e., rates of revascularisation by country), but this also did not account for the regional variation. This variable was not a strong predictor of mortality in HERO-2 as after adjustment for patient risk factors it was no longer significant as in Giugliano et al [52] who found no clear, simple relationship between revascularisation and short to medium-term

survival in InTIME-II. Conversely in another AMI trial, ASSENT-2, a significant relationship with 30-day mortality was demonstrated after adjustment for baseline variables [49]. However the overall rate of revascularisation was higher in ASSENT-2 (30% compared to 6.2 and 24% in HERO-2 and InTIME-II respectively). Most importantly, in that study adjustment for country-level revascularisation rates in multilevel modeling did not eliminate the inter-country variation, as was the case in HERO-2. In HERO-2 adjustment for country-level median length of hospital stay also did not explain the variation.

Domanski et al [80] was unable to explain the geographical variation in mortality in the MAGIC trial using baseline prognostic risk factors and individual treatments including concomitant drug use and attempted reperfusion.

Another potential contributing factor was the higher rates of enrolment of screened patients into the HERO-2 trial in non-Western regions, particularly in Eastern Europe and Russia where the randomisation rate was 3-fold higher compared to in Western countries. Consequently there were more high-risk patients in these regions on the trial as shown by the increased number of patients with Killip class ≥II. However the disparity in mortality rates across regions was already evident in the sample of patients included in the screening log substudy; in-hospital mortality was around 2-fold higher in non-Western regions despite Western patients being the eldest group. In this sample large variations in treatment were also apparent with much lower use of reperfusion therapy in Russia and Eastern Europe. Mortality was still markedly higher in Latin America and Asia compared to Western countries despite similar rates of reperfusion therapy. In the overall trial adjustment for differences in enrolment rates at the country level (i.e., % randomised of eligible patients in each country) did not help to explain the regional variation.

Regional differences in mortality were not due to dissimilarity in the completeness of follow-up as death status at 30 days was obtained for 99.9% of all patients. Regarding the major secondary endpoint reinfarction, for which Western countries had the highest incidence rate, regional differences were not due to a failure to identify events from the medical records as an independent audit showed that reporting of events was of a high standard across all regions. Also the same criteria needed to be met for a suspected reinfarction to be counted as a confirmed event and one independent committee adjudicated all suspected events. So it seems that the variation may be at least partially due to a difference in the detection of events; i.e., differences in sites monitoring practices such as performing enzyme and electrocardiogram checks.

The multilevel analysis confirmed the findings of the single-level analysis for regional effects and enabled a valid assessment of the association between factors measured at the country-level and 30-day mortality. The strength and statistical significance of associations between country-level variables and mortality was over-stated in the conventional logistic analysis consistent with Austin et al [50] who compared the two analytic strategies in patients hospitalised with AMI for the prediction of 1-year mortality. Our method of adjusting for patterns of care differed to that used in MAGIC and

InTIME-II where patient-level variables were used for therapies such as concomitant drug use, attempted reperfusion and revascularisation. The drawback of this approach is that these variables are not measured at baseline and patients must survive long enough to be administered these therapies. Hence we preferred to adjust for differences in the average use of revascularisation between countries. The alternate approach did not explain the regional variation in MAGIC or InTIME-II. Multilevel analyses were not attempted in these trial datasets.

In the multilevel modeling analysis none of the national health and economic variables explained the significant region effect or the country-level variation. The final inclusion of region in the model did explain the remaining variation due to unmeasured characteristics at the country level. This was contrary to the analysis in the ASSENT-2 data where country-level life expectancy was sufficient to explain virtually all of the remaining country-level variation [49]. In that analysis tests of significance were available as a specialised multilevel modeling software package was used; the country-level variance was highly significant ($P<0.001$) until life expectancy was adjusted for ($P>0.5$). However HERO-2 is substantially more geographically diverse; 85% compared to 7% of patients in ASSENT-2 were recruited in non-Western countries. In HERO-2 region is a proxy for factors pertinent to country not already included in the model.

The hospital-level variance in HERO-2 was never explained and was estimated at 0.063 in the final model. In ASSENT-2 the hospital-level variance was never significant and for the final model was very similar to in HERO-2 (i.e., 0.062; $P=0.287$). Our multilevel analysis had the limitation that we did not include physician- and hospital-level characteristics as this information was not available. A sensitivity analysis was done which included a variable on hospital volume constructed by classifying the participating hospitals into quartiles according to the number of patients enrolled; results were unchanged. However Austin et al [50] demonstrated that the level of unexplained variation attributable to variations between physicians and hospitals for 1-year mortality is small (2.8 and 0.6% respectively). Furthermore in that paper the effect of teaching hospital and on-site revascularisation capacity was not significant. The latter result accords with work by Llevadot et al [103] who demonstrated in InTIME-II that the availability of 24-hour on-site catheterisation facilities was not associated with 30-day and 1-year mortality outcomes.

This study has a number of limitations. The grouping of countries into regions, although pre-specified, was somewhat arbitrary, based on geographical, racial or ethnic similarities. Selective recruitment may have contributed to the observed regional differences. For instance the availability of on-site catheterisation facilities may have influenced the recruitment of patients onto the trial. Also an increased use of fibrin-specific thrombolytics (i.e., alteplase, reteplase etc) was noted among Western screened patients (data not shown) suggesting an increased availability of these more advanced treatments in Western countries. This may have also effected recruitment as the trial protocol stipulated thrombolytic treatment with streptokinase. Evidently in Western countries doctor or patient preference was more often cited as the reason for not randomising patients. Another limitation is that

mortality was higher among non-participants than participants and recruitment was higher in non-Western regions. However as stated earlier the regional variation was already evident among screened patients and differences in patient characteristics were adjusted for. Another issue concerns how representative were the hospitals that participated in the trial. Most likely participating hospitals included those that administer the best quality care. If this is so mortality may be even higher and more disproportionate.

Given the prior evidence the observed differences are unlikely to be due to the play of chance. Processes of care issues are highlighted by the shorter time to treatment in the countries with the greatest use of revascularisation (i.e., Western countries). In Eastern Europe, Russia and Asia the most common reason for not randomising was ineligibility due to late presentation (i.e., time since symptom onset >6 hours). We also conducted a landmark analysis including only patients who survived the first 24 hours (data not shown). The pattern of regional variation was consistent both before and after 24 hours. This suggests if differences in processes of care within the first 24 hours explain part of the variation then these differences persist after 24 hours. The longer hospital stays in Eastern Europe and Russia are questionable and indicate perhaps more cost-effective use of health resources is needed.

The large gap in survival rates from AMI between Western and non-Western regions needs to be closed. The cause of this discrepancy may be multi-faceted i.e., a combination of patient elements (e.g., lifestyle differences) and diversity in health care systems across regions. The results have elucidated differences in practices between regions and suggest that there is potential for improvement in the treatment of AMI in non-Western countries.

CHAPTER 5: RISK OF MORTALITY AFTER ACUTE MYOCARDIAL INFARCTION: PERFORMANCE OF MODEL UPDATING METHODS FOR APPLICATION IN DIFFERENT GEOGRAPHICAL REGIONS

5.1 Introduction

Randomised clinical trials follow rigorous scientific principles and are not subject to many biases inherent in other designs such as observational or retrospective studies [104]. Therefore they provide a rich source of information which, as required in the conduct of the RCT, is generally accurate and complete. Thus risk prediction models can be developed from such information for integration into clinical practice. However as mentioned earlier clinical trial-derived risk models may not generalise to real-world patients [58]. There are numerous potential reasons for limited generalisability; considerations include patient selection, geographical differences and temporal trends [105]. As a result a risk model may need to be updated for application to an external patient population.

Steyerberg et al [57] used the GUSTO-I trial as the basis for an extensive evaluation of methods for updating risk models for AMI in Western countries. Re-calibrations and other methods were applied in samples of study subjects from 7 US regions and 6 other groupings (Belgium, Netherlands–United Kingdom, middle Europe, Israel, Canada, Australia–New Zealand). *Simple re-calibration* methods (i.e., re-estimation of the intercept or both intercept and slope of the linear predictor) were found preferable to more extensive *model revision* methods (re-estimation of some or all regression coefficients, model extension with more predictors) which only performed better when applied to large datasets in combination with shrinkage.

This chapter compares methods for updating risk models in the HERO-2 trial. We extend the evaluation of Steyerberg et al [57] by applying methods to non-Western patient populations such as those of Russia, Eastern European, Latin American and some Asian countries. Simple re-calibration (re-estimation of the intercept and slope of the linear predictor within regions) and model revision (re-estimation of all regression coefficients within regions) with and without shrinkage are compared to risk assessed using a *global* additive model with built-in region adjustments (i.e., the HERO-2 reduced model derived in Chapter 3 which is relabeled here the HERO-2 global risk (HGR) model, since the coefficients, the β's, are common to all regions). The relative performance of methods in different geographical regions is of primary interest.

Shrinkage is a general technique to improve an estimator so that better calibrated predictions result when the model is applied elsewhere [25]. When a calibration plot (i.e., plot of observed versus predicted responses) is constructed from the same dataset used to fit the model parameters, the estimation process forces the calibration slope (i.e., slope of observed versus predicted values) to be one. However, when the model is applied to an independent dataset, over-fitting will cause the slope of the calibration plot to be less than one, a result of regression to the mean. The bias in predictive accuracy measures that results from over-fitting in the final model fit is referred to as optimism [25].

We applied shrinkage using penalised maximum likelihood estimation which executes the shrinkage during the estimation of model coefficients. This method also has the advantage that it can differentially shrink predictors and has been demonstrated to be less optimistic compared to heuristic and bootstrap shrinkage [106]. The method was applied to shrink estimated region-specific regression coefficients towards the global model coefficients.

Simple re-calibration methods are attractive because of their stability; however their disadvantage is a potential for bias in the individual regression coefficients. Model revision, an alternative approach to risk calculations for a new population is expected to lead to less bias but higher variance in the updated model coefficients and predictions. The aims of this work are to determine whether region-specific model revision provides worthwhile improvements in predictive accuracy over simple re-calibration. If so, interactions between the level of improvement and geographical region may exist. The extent of improvement will depend on both the data information available in a region and the extent of true heterogeneity in risk factor effects by region (interactions). We will also assess whether shrinkage provides any benefit in the smaller regions where it is likely to be most effective.

It is expected that, in the regions with large sample sizes, model revision with and without shrinkage will provide little benefit over simple re-calibration since these regions contribute much of the prognostic information to the global model. However, with Asia contributing little data to the trial, model revision in this region may improve prediction accuracy since it contributed less to the estimation of coefficients in the global model.

We deemed it appropriate to base this study on the reduced version of the HERO-2 model as opposed to the full model as it may be more common to use such a reduced model in practical application.

5.2 Methods

5.2.1 Model updating methods

We applied simple re-calibration involving region-specific estimation of the intercept and slope. A region-specific slope permits the estimate for the association between the LP of the logistic regression and outcome to vary by region. For this re-calibration, using the HGR model, we first formed, for every study participant, the LP that would be appropriate for them were they a Western patient, by subtracting the corresponding regional adjustment from the LP for subjects in non-Western regions. Designate the resulting "Westernised LP" to be LP_W. Next we fitted a logistic model within each region with this Westernised LP included as the only covariate i.e. $LP_R = \alpha_R + \beta_R LP_W$.

The incorporation of regional adjustments into the HGR model is, effectively, simple re-calibration via regional updating of the intercept. These constants are reproduced by fitting a logistic regression model containing Westernised LP as an offset and the variable region i.e. $LP_R = \alpha_R + LP_W$.

Model revision requires estimating region-specific coefficients. To do this, logistic regression models, containing the patient risk factors from the HGR model, were fitted for each region. Model revision

with additional shrinkage was applied using penalised maximum likelihood estimation in S-PLUS 7.0 using functions from the Hmisc and Design libraries [68]. Shrinkage of estimated region-specific regression coefficients towards the global model coefficients was achieved by fitting models to each region, including the HGR model LP as an offset term. The syntax and further explanation is shown in Appendix A.

5.2.2 Internal validation

With EPV ≥40, optimism in the apparent estimates of model performance is small [66]. Accordingly, the optimism in Russia and Eastern Europe should be negligible, but in the other regions, especially Asia, the apparent model performance estimates may be overestimated. To obtain optimism-corrected estimates of model performance, bootstrap resampling was conducted [73]. 200 bootstrap samples of size 17073 were drawn with replacement from the original sample. Samples were drawn with stratification by region so as to retain the original regional distribution. Models as estimated in the bootstrap sample (with simple re-calibration or model revision with and without shrinkage applied) were evaluated in the bootstrap sample (bootstrap performance) and in the original sample (test performance). For test performance the value of the LP was calculated for each patient in the original sample as the linear combination of the regression coefficients as estimated in the bootstrap sample with the values of the covariables in the original sample, and if applicable, with simple re-calibration applied using parameters (e.g. α_R, β_R) estimated from the bootstrap sample. The difference between bootstrap and test performances is averaged to obtain a stable estimate of the optimism in apparent performance, since bootstrap performance is apparent performance in each bootstrap sample. The mean optimism is then subtracted from the apparent performance (i.e., performance of models derived from and tested on the original sample) to give the *optimism-corrected estimate* of performance.

This procedure was followed for all but the calibration slope. The calibration slope, as estimated in each bootstrap sample, provides an apparent indication of miscalibration which could be overly optimistic. However, in the context of optimism and bootstrapping the calibration slope is the "uniform shrinkage factor" and the mean of the bootstrap estimates of test performance ('slopes') provide a stable and realistic estimate of the shrinkage factor. This uniform shrinkage factor, when multiplied by the estimated regression coefficients, provides predictions which calibrate better in future patients [107].

5.2.3 External validation

The performance of the HGR model (LP) with original weights and with the application of model revision (i.e., Western and Eastern European specific weights derived from the HERO-2 data) was assessed in the GUSTO-I trial dataset used in Chapter 3.

5.3 Results

The HGR model is displayed again in Table 5.1 for easy reference.

Table 5.1 Coefficients of the HERO-2 global risk model

Variable	Parameter Estimate (SE)
Intercept:	-3.385 (0.240)
Regional adjustment:	
Western countries	-
Latin America	0.641 (0.119)
Eastern Europe	0.402 (0.097)
Russia	0.452 (0.096)
Asia	1.018 (0.154)
Killip class*:	
I	-
II	0.740 (0.065)
III-IV	1.647 (0.097)
Slope of age:	
Killip class: I	0.080 (0.004)
II	0.052 (0.005)
III-IV	0.013 (0.008)
Systolic BP	-0.019 (0.001)
Heart rate†	0.032 (0.002)
Anterior STE	0.465 (0.055)
Female	0.445 (0.058)

*Killip class effects at age 65.

†For heart rate, values <70 are truncated at 70 and values >100 are truncated at 100.

5.3.1 Performance of HGR model and simple re-calibration

In Table 5.2 we show the apparent performance estimates of the HGR model and results obtained after correcting for the expected over-optimism yielded from the bootstrapping procedure.

The apparent discriminatory value of the HGR model when applied to the overall HERO-2 dataset was 0.813, as measured by the *c* statistic, reflecting excellent discriminatory ability. Presented with two patients the multivariable model would correctly rank these patients 81% of the time. The same high ability to discriminate applied similarly within each region, except in Asia where the discrimination was marginally less.

The calibration of model predictions, as indicated by the apparent calibration slope, was excellent in Western countries, Eastern Europe and Russia. In Asia and Latin America adjustments to the LP of about 11% in different directions were indicated.

Table 5.2 Bootstrap results for the HERO-2 global risk model

	Western Countries	Latin America	Eastern Europe	Russia	Asia
Apparent performance					
c statistic	0.807	0.823	0.798	0.821	0.759
R^2 (%)	20.2	28.6	22.5	29.1	18.1
Brier score	0.056	0.077	0.079	0.092	0.086
Calibration slope	1.04	1.11	0.96	1.01	0.89
Bootstrap performance*					
c statistic	0.809 ± 0.015	0.824 ± 0.015	0.799 ± 0.009	0.821 ± 0.007	0.761 ± 0.030
R^2 (%)	20.4 ± 2.2	28.6 ± 2.6	22.6 ± 1.4	29.1 ± 1.4	18.2 ± 4.6
Brier score	0.056 ± 0.003	0.076 ± 0.004	0.079 ± 0.003	0.092 ± 0.003	0.085 ± 0.007
Calibration slope	1.05 ± 0.08	1.11 ± 0.07	0.96 ± 0.03	1.01 ± 0.03	0.89 ± 0.12
Test performance*					
c statistic	0.807 ± 0.002	0.823 ± 0.001	0.798 ± 0.000	0.820 ± 0.001	0.758 ± 0.003
R^2 (%)	20.1 ± 0.3	28.5 ± 0.3	22.4 ± 0.1	28.9 ± 0.1	17.7 ± 0.6
Brier score	0.056 ± 0.000	0.077 ± 0.000	0.079 ± 0.000	0.092 ± 0.000	0.086 ± 0.001
Calibration slope	1.03 ± 0.02	1.10 ± 0.03	0.95 ± 0.02	1.01 ± 0.02	0.88 ± 0.03
Expected optimism*					
c statistic	0.002 ± 0.015	0.001 ± 0.015	0.001 ± 0.009	0.000 ± 0.007	0.003 ± 0.030
R^2 (%)	0.3 ± 2.2	0.2 ± 2.6	0.2 ± 1.4	0.1 ± 1.4	0.5 ± 4.7
Brier score	0.000 ± 0.003	0.001 ± 0.005	0.000 ± 0.003	0.000 ± 0.003	0.001 ± 0.007
Shrinkage factor	1.03 ± 0.02	1.10 ± 0.03	0.95 ± 0.02	1.01 ± 0.02	0.88 ± 0.03
Optimism-corrected performance					
c statistic	0.805	0.822	0.798	0.820	0.756
R^2 (%)	19.9	28.4	22.3	28.9	17.6
Brier score	0.057	0.077	0.079	0.092	0.087
Calibration slope	1.03	1.10	0.95	1.01	0.88

*Mean ± SE_B

Overall performance, as indicated by R^2 statistics, was highest in Latin America and Russia.

The optimism-corrected estimates of performance measures were very similar to the apparent estimates as the expected optimism estimates were negligible. The average calibration slope on test estimated in the bootstrapping procedure were very consistent with the apparent performance estimates of the calibration slope, and the standard deviation of test performance was no more than 3% in any region.

Applying **simple re-calibration via region-specific re-estimation of intercept and slope** is an alternative to the inclusion of regional indicators in the HGR model, and can be applied in other population groups. This approach gave an unchanged overall *c* statistic of 0.813 (0.812 after optimism-correction). By definition the *c* statistics for each region were unchanged. Calibration slopes obtained from the bootstrapping procedure were all very close to 1. Apparent measures of overall performance were again very similar to the HGR model results and optimism-corrected performance.

5.3.2 Performance of model revision

Table 5.3 shows the apparent performance (and bootstrapping results) for application of model revision without shrinkage. Compared to simple re-calibration improvement in the apparent discriminatory ability was most noted in Asia. Only slight improvement was evident in Western countries and Latin America and no improvement in Eastern Europe and Russia. Adjustment for over-optimism however lessened any apparently worthwhile gains from model revision. The overall apparent *c* statistic was 0.817 which after optimism-correction decreased to 0.812, no better than the performance of simple re-calibration. The same optimism occurred in the R^2 statistic and Brier score, which after optimism correction showed no benefit from model revision. In Asia a 5.5% absolute increase was noted in the apparent R^2 estimate, however the optimism-corrected estimate was markedly lower than that in simple re-calibration. In all regions except Asia the application of shrinkage produced virtually identical predictive performance results. In Asia the optimism-corrected

Table 5.3 Bootstrap results for model revision (without shrinkage)

	Western Countries	Latin America	Eastern Europe	Russia	Asia
Apparent performance					
c statistic	0.818	0.831	0.798	0.822	0.788
R^2 (%)	21.6	30.2	22.8	29.2	23.6
Brier score	0.056	0.075	0.079	0.092	0.081
Calibration slope	1	1	1	1	1
Bootstrap performance*					
c statistic	0.822 ± 0.013	0.833 ± 0.015	0.799 ± 0.008	0.823 ± 0.007	0.799 ± 0.028
R^2 (%)	22.5 ± 2.2	30.9 ± 3.0	23.0 ± 1.4	29.3 ± 1.4	25.9 ± 4.6
Brier score	0.055 ± 0.003	0.074 ± 0.005	0.079 ± 0.003	0.092 ± 0.003	0.079 ± 0.007
Calibration slope	1	1	1	1	1
Test performance*					
c statistic	0.814 ± 0.003	0.827 ± 0.003	0.797 ± 0.001	0.821 ± 0.001	0.776 ± 0.010
R^2 (%)	20.7 ± 0.6	29.2 ± 0.5	22.5 ± 0.1	29.0 ± 0.1	18.3 ± 7.5
Brier score	0.057 ± 0.000	0.076 ± 0.001	0.079 ± 0.000	0.092 ± 0.000	0.084 ± 0.001
Calibration slope	0.95 ± 0.07	0.96 ± 0.07	0.99 ± 0.04	0.99 ± 0.03	0.85 ± 0.17
Expected optimism*					
c statistic	0.008 ± 0.013	0.006 ± 0.016	0.002 ± 0.009	0.001 ± 0.007	0.024 ± 0.028
R^2 (%)	1.8 ± 2.4	1.7 ± 3.2	0.6 ± 1.4	0.4 ± 1.4	7.6 ± 9.1
Brier score	0.001 ± 0.003	0.002 ± 0.005	0.001 ± 0.003	0.000 ± 0.003	0.005 ± 0.007
Shrinkage factor	0.95 ± 0.07	0.96 ± 0.07	0.99 ± 0.04	0.99 ± 0.03	0.85 ± 0.17
Optimism-corrected performance					
c statistic	0.810	0.825	0.796	0.821	0.764
R^2 (%)	19.8	28.5	22.2	28.8	16.0
Brier score	0.057	0.078	0.079	0.092	0.086
Calibration slope	0.95	0.96	0.99	0.99	0.85

*Mean ± SE_B

c statistic (0.768) showed slight improvement from model revision with application of shrinkage and the R^2 statistic increased moderately (optimism-corrected $R^2 = 19.3\%$).

Figure 5.1 summarises bootstrap resampling validation results for Asia, comparing the various model updating strategies. The test dataset is the observed HERO-2 dataset. The figure shows that, whatever criterion is used, the optimism of model revision, particularly without shrinkage, provided a biased result suggesting a better performance than will be realised. (Compare median bootstrap with median test performance for (c) and (d).) Comparing solid lines, optimism-corrected performance, the application of shrinkage (d) more consistently provided a slight optimism-corrected benefit over the performance of the HGR model (a) and simple re-calibration (b).

Calibration slopes, for the test data, are shown in Figure 5.2 for all regions and model updating methods. The variability in estimates decreased with increasing regional sample sizes and was greatest in Asia. The performance of model revision improved marginally in Asia with the application of shrinkage (average calibration slope on test: (c) 0.85 vs. (d) 0.92).

In summary, for the sample as a whole, applying updating methods to the HGR model made little or no difference (Table 5.4). Overall results did not concur with the Asia region where model updating did affect predictive performance and in some instances was detrimental. A small advantage was obtained when model revision was applied with shrinkage.

Table 5.4 Updating methods considered for the HGR model and summary of performance measures (optimism-corrected) calculated for the overall HERO-2 dataset and only the Asia region

No.	Updating method	Parameters estimated	*c* statistic		R^2 (%)		Brier score	
			Overall	Asia	Overall	Asia	Overall	Asia
1	No adjustment: HGR model; $LP = \alpha_R+\beta_{1..9}$	0	0.812	0.756	25.8	17.6	0.080	0.087
2	Simple re-calibration: $LP = \alpha_R+\beta_R LP_W$	10 (5 × 2)	0.812	0.756	25.7	17.3	0.081	0.087
3	Model revision: $LP = \alpha_R+\beta_{R,1..9}$	50 (5 × 10)	0.812	0.764	25.6	16.0	0.081	0.086
4	Model revision + shrinkage: $LP = \alpha_R+\beta_{1..9}+\lambda_R+\gamma_{R,1..9}$	50 (5 × 10)	0.812	0.768	25.8	19.3	0.081	0.086

Subscript R denotes region-specific estimate. In method 2 LP_w is the linear predictor that would be appropriate for a Western patient. Method 4 involves estimating region-specific deviations (γ) from the HGR model coefficients with shrinkage applied. An adjustment to the intercept (λ) is required in order to have the overall mean of the shrunken model equal to the sample mean. See Methods Section 5.2.1 and Appendix A for further details.

5.3.3 External validation in GUSTO-I trial

On application to GUSTO-I data the HGR model performed well and calibrated better than regional estimation of coefficients (model revision); predicted mortality rates were 6.9 and 6.2%, respectively,

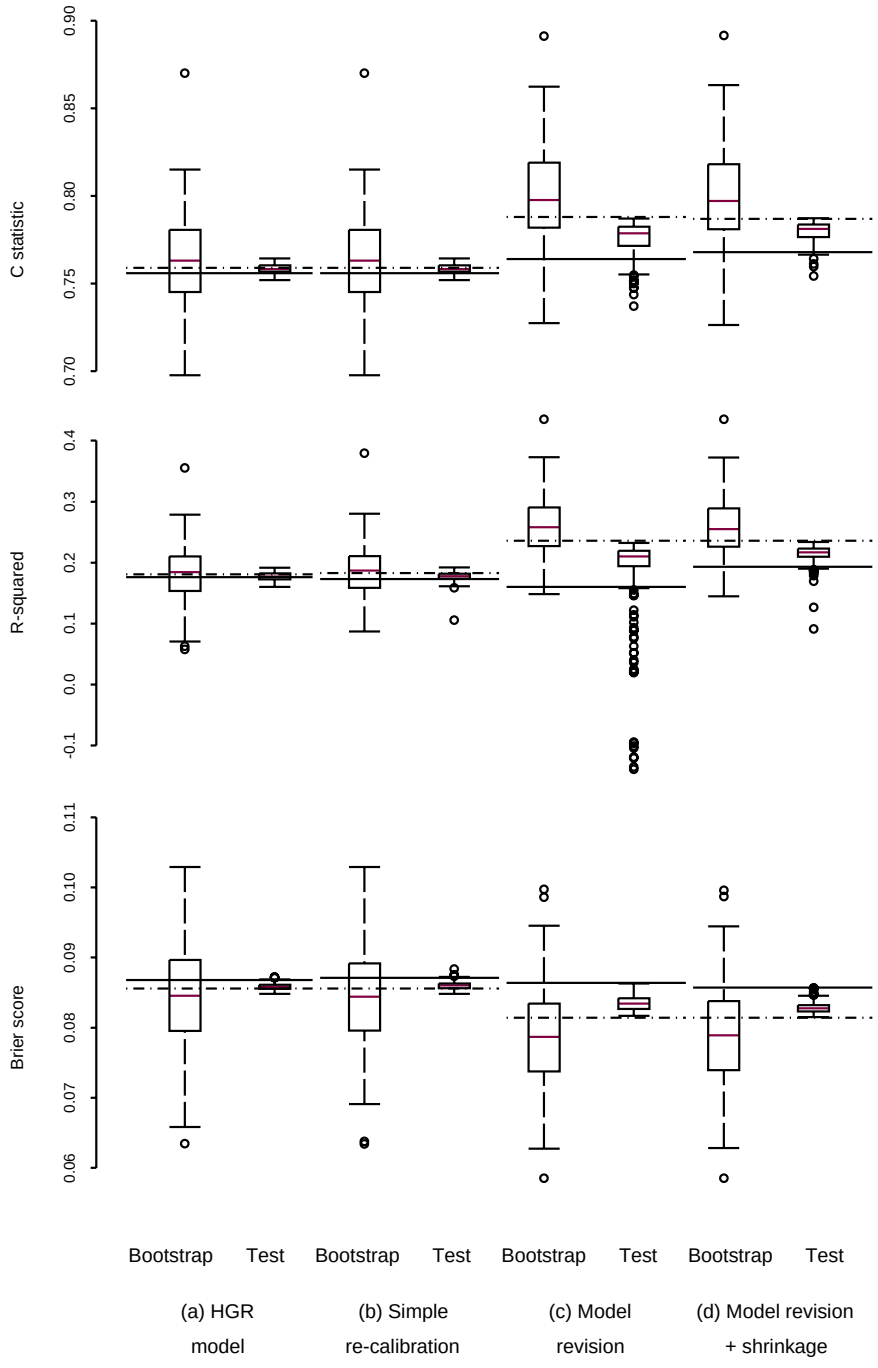

Figure 5.1 Box-plots are shown for the bootstrap resampling validation results for Asia, comparing the HGR model and the updating strategies. The dashed and solid lines indicate the apparent and optimism-corrected performance estimates respectively.

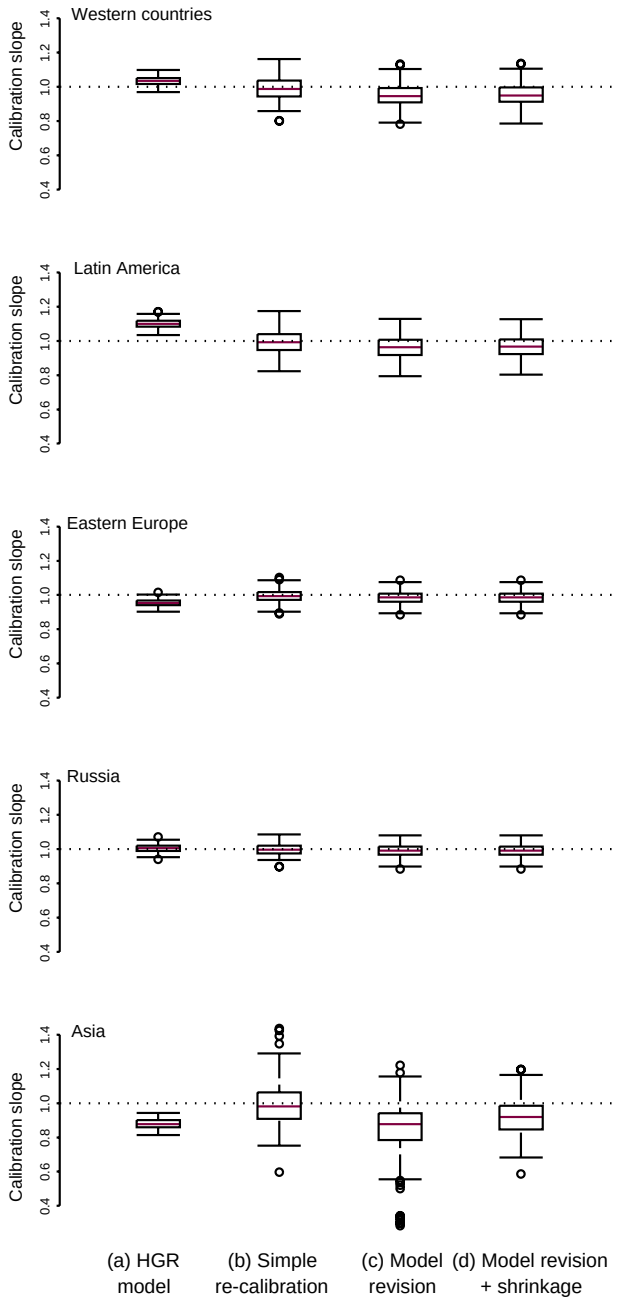

Figure 5.2 Box-plots are shown for test performance estimates of the calibration slope obtained in the bootstrap resampling validation procedure comparing the HGR model and updating strategies.

compared with the observed rate of 7.3%. Results for all performance measures are summarised in Table 5.5. The HGR model explained a marginally higher proportion of the variability in observed mortality (R^2).

Table 5.5 Application of the HGR model and region-specific coefficients to GUSTO-I

	HGR model	Region-specific model revision
c statistic	0.804	0.799
Brier score	0.059	0.060
R^2 (%)	21.5	19.9
Miscalibration:	$\alpha = 0.11, \beta = 1.02$	$\alpha = 0.02, \beta = 0.92$
P-value	0.04	<0.001

5.4 Discussion

Our risk estimates are derived from a trial based global model of 30-day mortality risk developed from international data. This model had excellent apparent discrimination overall, according to *c* statistics, and good agreement with data (R^2 and Brier scores) within the trial population. However, because of different sample sizes, some regions contributed more information than others. As expected, estimated over-optimism was negligible in the regions which were well represented.

In order to study the effects of model selection, we conducted an additional analysis where we applied backward selection to the full HERO-2 risk model in bootstrap samples. This strategy takes into account the uncertainty in the model selection process. In 71% of the bootstrap samples the model selected included the same risk factors that comprise the HGR model. This indicates that there was a model, consistently chosen, as might be anticipated given the high EPV -- at least 97 [26]. Given the consistency of this model selection, it is unlikely that an alternative modeling strategy for determining risk (for instance, Bayesian model averaging) would lead to much departure from HGR model predictions. Updating methods were then applied as before to the newly derived set of predictors in each bootstrap sample. All Tables and other results were virtually unchanged.

The important recent evaluation of model revision methodology of Steyerberg et al [57] suggests that re-calibration of an existing model for a new environment can be as effective as more complex statistical approaches. Steyerberg compared the performance of different model updating methods applied to a previously published model (TIMI-II, n=3339) in samples from the GUSTO-I population of over 40 000 patients ranging from small (n=200) to large (n=10 000). In their study, extensive model revision methods only performed better than simple re-calibration for samples of size at least 2000 and, even then, the gains achieved in discrimination and overall performance (i.e., Brier score) were very small. However, when the TIMI-II model was refitted in a smaller development sample (n=500), model revision did provide worthwhile gains for validation samples n≥500. Shrinkage

provided further improvement and was necessary to achieve a calibration slope near 1. This suggests model revision is more effective for models developed in small samples or specialised patient groups. This is plausible as the generalisability of these models would be questionable.

In the above paper, Steyerberg et al compared updating methods using a Western based model in patients chosen from Western countries. Our study extends this earlier study for application outside the West. We further found that estimation using region-specific coefficients (i.e., model revision) was unnecessary. Whether a model is generally applicable is a practically important question. The results here were consistent across regions; model revision did not provide worthwhile gains over simple re-calibration in both large and small regions and was often out-performed.

The lack of clinically significant interactions with region in the full HERO-2 model provides an important simplification of HGR model risk estimates. However, it should be noted that the sample size for the trial was determined to adequately power a comparison of treatment groups, and may not be large enough to provide power for tests of interactions between region and individual risk factors, particularly for rare conditions. Indeed, testing calibration slope may provide a more sensitive test of regional consistency in association of actual and predicted level of risk.

There were slight apparent improvements in predictive performance when model revision was applied in Asia. However due to small sample size most apparent gain was lost after adjustment for over-optimism, though shrinkage did retain a small advantage. The Asia region provided the most interesting results as patients from this region seemed the most unlike others. The majority of patients from this region were recruited in India (n=565) and Malaysia (n=108). Only 62 patients were recruited from countries with Western-style health care systems, such as Singapore and Hong Kong.

A more diverse study population with more events may have provided better information concerning risk factors thus potentially affording the HERO-2 risk models enhanced generalisability. However generalisability of risk factor effects may be hampered in the following situations. Temporal changes in the effects of predictors may occur as a result of new treatments emerging. For example the benefit of a new treatment may be dependent on certain patient characteristics. Differences in the definitions of predictors or in correlations between predictors in the development sample and the new setting may mean that some individual coefficients no longer apply [57]. And finally, true heterogeneity in effects of predictors may exist because a risk factor may not be measured.

Our study differed to that of Steyerberg et al in two major respects. Firstly we did not update a model developed in a previous study. This is because no globally based risk models for 30-day AMI mortality, up to now, exist. Further an important question is whether the accuracy of globally based risk model predictions can be further improved by estimating region-specific coefficients. Secondly, in internal validation we did not test updated models on independent data by partitioning regions into separate training and test set portions. We preferred to utilise all available data to estimate updated models as reliably as possible. Bootstrapping was employed to obtain optimism estimates to correct

apparent performance measures. As anticipated, given the large sample size and event numbers, in most regions, the over-optimism was small.

In Chapter 3 we examined various predictive models in HERO-2 and demonstrated that simplification resulted in little loss in predictive performance across all geographical regions. Therefore we expect that the results obtained here, in this investigation of updating strategies for the HGR model, would also apply to the full HERO-2 risk model.

In this study simple re-calibration in regions by updating of the intercept has no effect since the HGR model already involves region. Simple re-calibration is more generally applicable when the target population is different, or when there is the possibility of interactions depending on levels of risk in the population, where re-calibration of slope is informative.

What is the best basis for risk prediction? Augustin et al [26] compared approaches such as Bayesian and bootstrap model averaging to commonly used methods. These included risk prediction based on the full model (where all available covariates were included) and a single model selected using backward elimination (as applied here). They found that with sufficient events per variable (i.e., EPV ≥10) the model averaging approaches did not improve predictions. However the Bayesian and bootstrap model averaging procedures provide posterior probabilities and inclusion frequencies, respectively, for both models and variables and therefore are useful for obtaining a better understanding of the data and for judging the certainty attached to a selected model. The full model, whilst it ensures that all relevant prognostic information is captured, has the disadvantage that it requires information on more risk factors; this may present difficulty if the model is applied in another study. As well, for models developed in small studies, parsimonious models may be more reliable. Sample size should be a major consideration when selecting a model development strategy. The decision of whether to apply a comprehensive model or a reduced form will depend on the setting and ease of application. However the potential for systematic over- or under-estimation resulting from omitting practically significant risk factors should not be overlooked.

The point might be made that a model is useful for a new country, or for a new population of patients within a country, if and only if it does not need re-calibration, that is, is stable across a range of countries. The model itself should explain the inter-country differences - in principle at least - not just represent them by geographical parameters. More commonly, models developed for use in prediction of AMI in new populations will not be effective without re-calibration. In Chapter 4 we attempted to explain international differences in mortality rates using patient risk factors, treatments and economic statistics. The variation in mortality and other clinical outcomes across geographic regions was not adequately explained. Differences in the characteristics of patients not captured through measurement of traditional risk factors may be part of the explanation of the need for regional adjustment. For example genetic factors, ethnic origin or lifestyle may contribute, but were not measured in the trial. We prefer to include in risk models only risk factors that prove themselves in stratified analysis - i.e., with proven association to mortality within regions - not proxy factors that correlate with such

unknown characteristics of the population. Simple re-calibration with proven risk factors then remains appropriate in new populations.

Of further interest is whether risk models for acute ST-elevation MI can be successfully re-calibrated to predict outcomes for patients with other acute coronary syndromes (e.g., UA or non-Q-wave AMI). An alternative would be more extensive model revision, with consideration of additional variables. Granger et al [12] derived a simple model to assess risk for in-hospital mortality for the entire spectrum of ACS using a global registry. External validation was performed using the GUSTO-IIb dataset of 12 142 patients with the full spectrum of ACS; the model performed well in both the ST-segment and the non-ST-segment elevation subgroups. Furthermore the model identified risk factors which were similar to those of other AMI mortality models suggesting that simple re-calibration of existing models is appropriate for evaluation in patient populations with other acute coronary syndromes. This is however beyond the scope of this work.

In conclusion, this chapter supports the findings of Steyerberg et al [57] that extensive model revision should only be attempted with relatively large validation samples in combination with shrinkage. Simple re-calibration in each new setting, requiring local data collection, is otherwise appropriate. At present model revision is only indicated for models developed on small samples or specialised patient groups.

CHAPTER 6: CROSS-TRIAL COMPARISONS OF RISK MODELS

6.1 Introduction

Virtual Coordinating Centre for **G**lobal C**o**llaborative Cardiovasc**u**lar **R**esearch (VIGOUR) is a global collaboration of coordinating centres and investigators experienced in the conduct of large cardiovascular clinical trials [108]. It developed from a collaboration of investigators participating in the large, international Global Utilisation of Streptokinase and tPA for Occluded Coronary Arteries (GUSTO-I) trial, which began in 1990 [37]. The VIGOUR collaboration has a number of interests some of which include: examining mechanistic insights into how new diagnostic and therapeutic strategies exert their effects, examining processes of care and effects on patient outcomes, and evaluating international differences and practice patterns and their effects on outcomes. VIGOUR has accumulated large databases from serial trials to address key temporal trends, overviews, and subsidiary questions of importance to patients, care providers, sponsors, and regulators.

Through the Clinical Trial Centre's involvement in VIGOUR we have acquired data for four additional trials involving AMI patients. These trials include ASSENT-2 [60], ASSENT-3 [38], GUSTO-2b [61] and GUSTO-3 [62]. For this chapter we will use these trials, together with HERO-2 and GUSTO-I, to readdress the issue of simple-calibration versus model revision by adapting and modifying approaches taken in earlier chapters, particularly Chapter 5. This additional data brings the combined database to almost 90,000 patients and will allow some additional questions of interest to be examined. Before going further we recap earlier material which is relevant to this chapter.

In Chapter 3 we developed the HERO-2 model using a geographically diverse population and found that it performed well when applied to the GUSTO-I trial a Western patient population. However when we applied the GUSTO-I model to the HERO-2 heparin treatment arm we found that it had lower discriminatory ability in Asia and underestimated risk in non-Western patients. In Chapter 4 we explored differences in mortality rates between regions and found that there were no risk covariates available that fully explain the regional variation. This means that existing models developed in one region exhibit some bias, with resulting lack of accuracy, when applied to patients in other regions. This is a caveat in the recommendation for existing models, necessitating the need for model updating [57]. In Chapter 5 we conducted a study which compared simple calibration with model revision when both approaches are applied to a fitted reduced model, within regions of the HERO-2 dataset. This analysis investigated whether a global model or a model introducing region-specific weights was preferable. A limitation of this study was that independent data was not used to compare performance, so there was limited gain to be expected in regions which contributed most of the data.

In this chapter we will continue with the comparison of simple calibration versus model revision, however we will consider it at the *trial level*. The question to be addressed is whether new models need to be developed from scratch, for a specific trial, or whether existing models with or without recalibration, can be applied. Chapter 3 touched on this issue, but this chapter will conduct a more

thorough investigation. We will compare the performance of models in their derivation datasets with the performance on the same data of models developed in other trials. Additional trials also enable us to validate and compare existing models on entirely independent data. In this instance, no model will have the advantage of being derived on the test dataset. The analysis will provide further validation of existing models.

Models compared will include the HERO-2, GUSTO-I [1] and ASSENT-3 [9] models. As described earlier, the ASSENT-3 model is a dynamic risk assessment model which incorporates clinically relevant information unfolding during the hospital stay allowing risk assessment over the following forecasting periods beginning at day 0 (baseline), 3 h, day 2 and day 5. We will only consider the baseline model which incorporates variables based on electrocardiographic data. Such data are not commonly included in trial databases since they are complicated to summarise. Variables included ST-segment deviation and QRS score [109]. Comparison of this model with the others, which contain more traditional and simple patient risk factor information, will establish whether or not it is worthwhile incorporating such variables. Since these variables are not available in the other trials, comparisons with the ASSENT-3 model will be limited to its derivation dataset with estimates of apparent performance obtained for each model.

Since no model incorporating only simple risk factor information has been developed using the ASSENT data we will also derive a new model based on the ASSENT-2 trial data. The same candidate variables considered for inclusion in the HERO-2 model and modeling approach will be used to develop this model. In effect, this is application of a more extensive version of model revision [57] to the HERO-2 model in the ASSENT-2 trial.

In addition we will examine how models perform when applied to later related trials and whether there is an advantage. Since similar sites participate in serial trials, patients are more similar. For example, does the ASSENT-2 model have an advantage over the HERO-2 and GUSTO-I models when applied to the ASSENT-3 trial population? In relevant work König et al [110] compared the performance of validation approaches using a logistic regression model developed to predict functional independence in stroke patients. Internal validation approaches including ten-fold cross-validation (CV) and leave-one-clinic-out CV were employed to determine how well they estimate the temporal and geographical transportability of the model. For temporal validation, models are applied to patients recruited in the same clinics but during a later time period. Geographical transportability, also referred to as external validation, involves applying models to patients recruited from different clinics to those used for model building. Accuracy in actual temporal validation data was well predicted from ten-fold CV; however when predicting geographical transportability all approaches had difficulties. In temporal validation the model was not well calibrated, however neither shrinkage nor inflation of regression estimates was indicated. However in external validation calibration was low and predictions were too extreme indicating that both re-calibration and shrinkage was required.

Consequently we anticipate that calibration may be an issue when applying models to later related trials, due to temporal differences. However since similar hospitals take part in serial trials, patients will be similar, and therefore we expect the relationship between risk factors and death, projected by models, to hold.

However when applying models to other geographical regions, hence ensuring different hospitals, miscalibration involving the need for shrinkage or inflation of model coefficients will be of interest. Decreases in discrimination, suggesting model revision (i.e., re-estimation of regression coefficients) may be of benefit, will be noted. The difference here compared with Chapter 5 is that a model developed in one or more specific regions will be assessed in another separate region, as opposed to comparing the performance in each region of a global model and a model with coefficients estimated in that region.

After the application of simple calibration to a risk score that risk score may prove as good as the full model when applied in other populations. It is worth knowing, in selected validation datasets, the loss from using the HERO-2 risk score (HPI) in place of the full model. Comparisons will also be made with the previously developed TIMI risk score [3].

To summarise the aims are:

- To determine whether new models need to be developed for new trials or whether recalibration of existing models is sufficient. Comparison of the performance of the HERO-2, GUSTO-I and ASSENT-2 models in respective derivation datasets will address this.
- To compare the performance of models in entirely independent data. This will also provide further external validation.
- To develop a new model based on the ASSENT-2 data.
- To compare the performance of the ASSENT-3 model to models only containing traditional risk factors (i.e., no ECG indices).
- To determine whether models display an advantage when applied in serial trials.
- To gauge temporal and geographical validity – including assessment of whether the results concur with König *et al* [110] and the earlier chapters.
- To examine the relative performance of the HERO-2 full model and risk score outside of HERO-2.
- To compare the performance of the HPI and TIMI risk scores in independent data.

6.2 Methods

6.2.1 Trials and included patients

Table 6.1 lists all VIGOUR trials with the six trials included in this chapter bolded. For these six trials patients came from 55 countries and have been grouped into the same geographical regions used in HERO-2. There were three additional countries that did not participate in HERO-2 for which discussion was required regarding the appropriate classification. These included Iceland, Israel and the

United Arab Emirates which were categorised as Western countries on the basis of their economic standing.

Table 6.1 VIGOUR trials: trials included in this chapter are bolded

Trial	Sites	Patients	Countries	Description
ASSENT-II	**1021**	**16,949**	**29**	**Tenecteplase vs. alteplase for acute MI**
ASSENT-III	**575**	**6116**	**26**	**Tenecteplase/IV heparin ± abciximab vs. tenecteplase/enoxaparin for acute MI**
ASSENT-III Plus	88	1639	12	Tenecteplase + IV heparin or enoxaparin in prehospital setting for acute MI
GUSTO-I	**1081**	**41,021**	**15**	**Four thrombolytic strategies for acute MI**
GUSTO-IIa	227	2564	12	Hirudin vs. heparin in ACS
GUSTO-IIb	**373**	**12,142**	**13**	**Hirudin vs. heparin in ACS**
GUSTO-III	**807**	**15,060**	**15**	**Reteplase vs. alteplase for acute MI**
GUSTO-IV ACS	458	7825	24	Abciximab vs. Placebo for ACS
GUSTO-V AMI	820	16,588	20	Reteplase ± abciximab for acute MI
HERO-II	**539**	**17,073**	**46**	**Bivalirudin + Streptokinase vs. IV heparin + Streptokinase for acute MI**
PARAGON-A	279	2282	21	Lamifiban vs. heparin in ACS
PARAGON-B	388	5225	30	Lamifiban vs. heparin in ACS
SYMPHONY	670	9233	33	Sibrafiban vs. aspirin, secondary prevention in ACS
2nd SYMPHONY	625	6671	35	Sibrafiban ± aspirin vs. aspirin alone, secondary prevention in ACS

All of these trials compared therapies for the treatment of suspected MI with the exception of GUSTO-2b which included patients with all types of acute coronary syndromes. In the GUSTO-2b clinical trial report patients were stratified according to the presence (n=4131) or absence (n=8011) of ST-segment elevation on the baseline electrocardiogram and considered to have non-Q-wave MI or unstable angina respectively. Analysis of all endpoints was stratified this way. For the sake of simplicity and since the subject matter here is risk assessment models for MI, only patients in the non-Q-wave MI stratum of GUSTO-2b are included in this work.

For analysis purposes patients were classified according to whether they received streptokinase or the more modern fibrinolytic therapies known as tissue plasminogen activators. These include drugs such as alteplase, reteplase and tenecteplase which have proven to provide additional survival benefit compared to streptokinase [37].

The GUSTO-I trial compared four treatment strategies: streptokinase and subcutaneous heparin, streptokinase and intravenous (IV) heparin, accelerated t-PA and IV heparin, and both t-PA and streptokinase with IV heparin. To ease the application of other models to the GUSTO-I trial, the t-PA + streptokinase combination treatment arm has been excluded from this chapter.

In all trials except GUSTO-I randomised treatment arms were pooled as no significant difference between treatments for their effect on 30-day mortality was demonstrated.

6.2.2 ASSENT-2 model development

The model development procedure used to derive the HERO-2 model was followed for the ASSENT-2 model. The same candidate predictors were considered with the caveat that the predictor variables **angina** and **history of cerebrovascular disease** were not available as the ASSENT trials did not collect this information. To reiterate: for the 16 949 patients enrolled in ASSENT-2, generalised additive models were applied in both univariate and multivariable contexts, to assess the shape and strength of the relationship between continuous clinical variables and 30-day mortality. Consequently **systolic BP** was fitted with a non-linear transformation applied; values above 140 were truncated at 140, since the spline relationship between **death** and **SBP** was constant for these values. Backward stepwise logistic regression was performed; sixteen baseline variables entered the initial model and were retained if $P<0.05$. Interactions were assessed for variables that remained in the model; only significant interactions which were also deemed clinically important were retained. The variable **region** was considered for inclusion in the model after independent predictor variables, and interactions between them, had been identified. Interactions between **region** and the baseline characteristic variables were assessed starting with a global test of additivity [25]. The derived model is referred to as the ASSENT-2 model.

6.2.3 Application of risk models

When applying the HERO-2 full model to the ASSENT trials the effect of **angina** was not taken into account since this variable was not available. This means that the linear predictor was calculated as the linear combination of the regression coefficients for the HERO-2 full model and values of the covariables for each patient in the ASSENT datasets ignoring the **angina** variable. However given that in any model the magnitude of explanatory variable coefficients depend on the presence of other correlated variables in the model [25], simply applying the HERO-2 full model ignoring the **angina** term may result in miscalibration. To gauge the extent of this dependency in this context we refitted the HERO-2 full model to the HERO-2 dataset, excluding **angina,** to investigate whether the parameter estimates for the other predictors change when the effect of **angina** is not adjusted for.

Most of the resulting model coefficients were very similar to the HERO-2 full model. However there were some differences which may not seem large but have an influence. The **prior MI** coefficient increased from 0.249 to 0.320, the **hypertension** coefficient increased from 0.130 to 0.164, and the **region** adjustment for Russia increased from 0.279 to 0.345. For the continuous variables the parameter estimate for **age** in **Killip class** I increased from 0.0772 to 0.0776; in the higher Killip classes even larger increases in the effect of **age** were observed (data not shown). The **heart rate** parameter estimate increased from 0.0301 to 0.0303. The increased effects for these continuous variables seem small but make a difference as they affect all patients.

Consequently both the original HERO-2 full model (ignoring the **angina** term) and this alternate model, which excludes **angina**, were applied to the ASSENT trials. This alternate model was also applied to the other trials to compare its performance with the complete HERO-2 model when information on **angina** is available.

In this chapter the GUSTO-I model will be applied to both the heparin and bivalirudin treatment arms of the HERO-2 dataset; in Chapter 3 the bivalirudin group was excluded because its treatment term does not allow for bivalirudin. However there was no difference in 30-day mortality according to randomised treatment in the HERO-2 trial, so we have chosen to extend the inclusion criteria here. Accordingly the risk coefficient for the streptokinase + IV heparin treatment term in the GUSTO-I model will be applied to bivalirudin patients.

The risk factor **history of cerebrovascular disease** will not be taken into account when applying the GUSTO-I model to ASSENT trial patients; as stated earlier this information was not collected. The resulting bias would be negligible as the typical proportion of patients with this risk factor is very small (<2.5%) [1, 39, 61-62].

6.2.4 Missing data

Trials exhibited low rates of missing data on all variables other than **time from symptom onset** (17% missing in GUSTO-I), **height** (6 – 12% missing in all trials except HERO-2 (2%)) and **location of MI** (11% missing in GUSTO-2b). For all other variables missing rates were less than 3%. In GUSTO-I, where simple imputation was applied throughout, other variables with missing value rates above 1% included **pulse** (2.2%), **SYSBP** (2.0%) and **weight** (2.4%). The rules and methods applied for imputing missing data are described in Methods Section 2.1.2.

Rates of missing data for 30-day mortality were very low across all trials (<0.5%) and all patients with unknown mortality status were assumed alive.

One unusual case was removed from the GUSTO-2b dataset; this patient was characterised by a very large outlier for **time from symptom onset to randomisation** (>1 month).

6.2.5 Treatment offset

To assess how well models perform without recalibration, when applied to other trials, it is necessary to adjust for differences in the type of fibrinolytic therapy used. These adjustments can be made by using an offset term [81]. An offset can be thought of as a variable that is included in the model but whose coefficient, $\hat{\beta}$, is not estimated but rather is forced to be equal to 1. When calculating LP scores including an offset term a predetermined amount is added to the linear predictor.

As stated earlier in GUSTO-I t-PA therapy was demonstrated to provide additional survival benefit over streptokinase. In other trials treatment effects were absent. The HERO-2 trial used streptokinase for fibrinolytic therapy, so to apply the HERO-2 model to trials using t-PA therapies an offset needs to be incorporated otherwise risk predictions will be over-estimated. Based on the results of the GUSTO-

I trial a risk reduction of 14% was assumed for the effect of therapy with t-PA relative to streptokinase. This means that 0.151 is subtracted from the calculated LP when applying the HERO-2 models to patients treated with t-PA therapies. Conversely if applying the ASSENT-2 model to HERO-2 patients 0.151 is added to the LP score since the ASSENT-2 patients were administered either tenecteplase or alteplase (i.e., t-PA).

Patients in the GUSTO-2b trial with ST-segment elevation received t-PA (52%), streptokinase (22%), or neither, at the attending physician's discretion for fibrinolytic therapy. Based on the GISSI trial [111] an 18% risk reduction was assumed for the effect of streptokinase versus no fibrinolytic therapy. So for patients treated with no fibrinolytic therapy values of either 0.349 or 0.199 were added to the model LP, which correspond to increases of 42% and 22% in the odds of death relative to patients treated with t-PA and streptokinase respectively.

The GUSTO-I model incorporates adjustment for treatment with t-PA or streptokinase, so further adjustment of the calculated LP was only necessary for GUSTO-2b patients treated with no fibrinolytic therapy. The adjustment for no fibrinolytic therapy versus t-PA described above was applied, as the reference group for the GUSTO-I model treatment term is accelerated t-PA and IV heparin.

6.2.6 Approach to chapter objectives

6.2.6.1 Comparison of trials

Firstly the characteristics of the trials included in this chapter will be documented. Then patients will be compared between trials. This will involve examining geographical, demographic and clinical characteristics. To assess differences in the overall risk profiles of patients, prognostic indices will be used including the HPI and TIMI risk scores. Following this 30-day mortality rates will be compared between trials both overall and stratified by region.

6.2.6.2 Comparison of models

Models will be compared by considering selected predictors, coefficient values and the relative amount of prognostic information contributed by predictors. The last listed aspect will employ the global χ^2 statistic from the logistic model and a comparison of this index from the full model containing all variables with a reduced model excluding the predictor of interest [81]. The difference indicates the independent contribution of each variable after adjustment for all other factors and demonstrates the importance of a given predictor. A likelihood ratio test can be performed to obtain a P-value.

6.2.6.3 Re-calibration

Models will be applied as they are, i.e., without re-calibration, to whole trials unrestricted by region and within regions. Subsequently *trial-level* re-calibration will be applied by fitting the following model: logit(observed mortality) ~ α_{TRIAL} +β_{TRIAL}(LP+offset). LP refers to the model linear predictor

as described in Chapter 3 and offset is the adjustment for fibrinolytic treatment described above. The performance of trial-level calibration will be assessed for trials overall and subdivided by region.

In addition *trial-region* level re-calibration will be applied. This will involve estimating trial-region specific intercepts and trial specific slopes according to the following model: logit(observed mortality) ~ $\alpha_{TRIAL,\ REGION}$ +β_{TRIAL}(LP+offset). Hence within each trial, region specific intercepts will be estimated which will correct for regional differences but a common slope will be assumed for the relationship between the LP and mortality. The effect on R^2 of this additional level of calibration will be of main interest.

6.2.6.4 Calibration of prognostic indices and comparison with HPI-FULL

Prognostic indices including the HPI and TIMI risk scores (refer to Chapter 3 for details on included risk factors and weights) will be calibrated by trial and region by fitting the following model: logit(observed mortality) ~ $\alpha_{TRIAL,\ REGION}$+(β_{TRIAL}Risk score)+offset. The offset will only be applicable in GUSTO-I as this was the only trial to demonstrate a significant effect of fibrinolytic treatment; i.e., the offset constant will be zero in the remaining trials since adjustment for treatment is not necessary. This is equivalent to including a term for treatment in the model except the effect of treatment is forced based on published results (described above in Section 6.2.5) as opposed to re-estimating it from the data.

To facilitate a fair comparison of the risk scores with a risk model, trial-region level re-calibration will be applied to the HERO-2 full model in this comparison.

The predictive performance of these risk assessment strategies will then be assessed in trials overall and subdivided by region. The need for region specific slopes when applying simple calibration to risk scores is also considered.

6.2.6.5 Predictive performance

Predictive performance measures will be calculated for each trial both overall and within regions.

The application of trial-level re-calibration by definition will not change *c* statistics. For trial-region level re-calibration overall results may change.

To assess calibration we used the intercept and slope framework as described in Methods Section 2.2. To reiterate the following logistic model is fitted: observed mortality ~ α +β linear predictor. Well calibrated models have an intercept α of zero and a slope β of one meaning no adjustment to risk forecasts is required. After recalibration is applied overall results will not be useful as by definition the intercept will be zero and the slope one. Results within regions will still provide information. After trial-region level re-calibration the calibration slope will only be useful as by definition calibration-in-the-large will be perfect overall and within regions.

Because in this approach the estimate of the intercept depends on the estimated slope, the adjustment to patient risk levels to correct calibration-in-the-large is only uniform when the slope is 1.

6.3 RESULTS

6.3.1 Trial characteristics

Characteristics of the trials included in this chapter are described in Appendix B Table A1. The GUSTO trials ran from 1990 through to 1997 and were followed by ASSENT-2, HERO-2 and then ASSENT-3 which began recruiting in May 2000. All of these trials were multinational and compared either fibrinolytic or antithrombotic therapies for the treatment of AMI or unstable angina.

The eligibility and inclusion criteria for the 6 trials were very similar except for GUSTO-2b which enrolled patients with chest discomfort within 12 hrs as opposed to the 6 hour cut-off used in other trials. Also less severe symptoms needed to be exhibited on the baseline electrocardiogram since the target population was all ACS and not exclusively AMI.

All trials had 30-day mortality as their primary outcome or at least as a component if this were a composite endpoint.

Only the GUSTO-I trial established a significant difference according to randomised treatment for 30-day mortality. This trial found that the combination of accelerated t-PA with heparin provided a survival benefit over previous standard thrombolytic regimens including streptokinase administered with either intravenous or subcutaneous heparin.

6.3.2 Patient characteristics by trial

As already discussed the representation of patients from non-Western countries in HERO-2 was high. In comparison the ASSENT and GUSTO trials recruited predominantly from sites in Western countries (Table 6.2).

Table 6.2 Regional distribution of VIGOUR patients

		Region, n (%)				
Trial	**Overall, N**	**West**	**EE**	**Russia**	**LA**	**Asia**
GUSTO-1†	30647	30375 (99)	272 (1)	-	-	-
GUSTO-2b‡	4130	4130 (100)	-	-	-	-
GUSTO-3	15059	14052 (93)	886 (6)	-	121 (1)	-
HERO-2	17073	2563 (15)	5877 (34)	6057 (36)	1820 (11)	756 (4)
ASSENT-2	16949	15754 (93)	475 (3)	-	720 (4)	-
ASSENT-3	6095	5465 (90)	90 (1)	-	540 (9)	-

†Excludes t-PA + streptokinase + IV heparin treatment arm; ‡ST-segment elevation stratum only.
West = Western countries; EE = Eastern Europe; LA = Latin America.

Table 6.3 documents patient demographic and clinical characteristics. There were higher proportions of **females**, patients with **hypertension** and patients classified as **Killip class** II or higher in HERO-2. The proportion of patients in HERO-2 having a **previous revascularisation** was smaller than in the

other trials. There was substantial variation in the delay from **symptom onset to randomisation** across the trials; the proportion of patients receiving treatment more than 4 hours after symptom onset varied from 12% in GUSTO-I to 35% for the AMI stratum in GUSTO-2b.

Table 6.3 Characteristics of VIGOUR patients, by trial*. Numbers are percentages unless indicated otherwise.

		Demographics:				
Trial:	**N**	**Age (yrs)§**	**Age 75+ yrs**	**Female**	**Weight (kg)§**	**Height (cm)§**
GUSTO-1†	30647	62 (52, 70)	12	25	78 (70-89)	172 (165-178)
GUSTO-2b‡	4130	63 (53, 71)	15	24	77 (68-86)	170 (165-177)
GUSTO-3	15059	62 (52, 71)	16	27	78 (70-87)	171 (165-177)
HERO-2	17073	62 (52, 70)	13	28	75 (68-85)	170 (164-175)
ASSENT-2	16949	61 (52, 70)	15	23	78 (69-87)	170 (165-177)
ASSENT-3	6095	61 (52, 70)	14	24	78 (69-88)	170 (165-176)

	Risk Factors:			Cardiovascular history:				Treatment:
	Current smoker	**Diabetes**	**HTN**	**Angina**	**Prior MI**	**Prior CVD**	**Prior Revasc**	**symptom onset >4 hours**
GUSTO-1†	43	15	38	37	17	2.0	8	12
GUSTO-2b‡	41	16	40	48	17	1.7	10	30
GUSTO-3	41	16	39	41	18	2.5	8	18
HERO-2	44	14	52	47	15	1.3	2	29
ASSENT-2	45	16	38	-	16	-	8	19
ASSENT-3	46	18	41	-	14	-	8	19

	Presenting characteristics:					
	Anterior MI	**SYSBP (mm Hg)§**	**DIABP (mm Hg)§**	**Heart rate (bpm)§**	**Killip class:**	
					II	**III or IV**
GUSTO-1†	39	130 (112-144)	80 (70-90)	73 (62-86)	13	2.1
GUSTO-2b‡	44	130 (115-148)	80 (70-90)	74 (64-86)	10	1.6
GUSTO-3	48	135 (119-150)	80 (70-90)	73 (62-86)	12	1.9
HERO-2	45	134 (120-150)	80 (70-90)	76 (65-88)	18	3.9
ASSENT-2	40	133 (120-150)	80 (70-90)	72 (62-85)	10	1.6
ASSENT-3	39	132 (119-150)	80 (70-90)	73 (62-85)	10	1.5

**P*-value<0.001 for all comparisons by trial. †Excludes t-PA + streptokinase + IV heparin treatment arm. ‡ST-segment elevation stratum only. §Median (IQR). CVD = cerebrovascular disease; Revasc = revascularisation; HTN = hypertension; DIABP = diastolic blood pressure.

Overall risk, measured with the HPI and TIMI risk scores, is summarised in Table 6.4. Patients in the ASSENT trials were associated with lower average risk profiles. HERO-2 patients had the highest overall risk however this is because of a high representation of Russian patients who had higher average risk profiles.

Table 6.4 Risk scores by trial and region

		Region					
Trial	**Overall**	**West**	**EE**	**Russia**	**LA**	**Asia**	***P*‡**
HPI risk score				***Mean (SD)***			
GUSTO-1*	2.7 (2.6)	2.8 (2.6)	2.4 (2.5)	-		-	0.021
GUSTO-2b†	2.8 (2.5)	2.8 (2.5)	-	-		-	-
GUSTO-3	3.0 (2.6)	3.0 (2.6)	2.9 (2.6)	-	2.8 (2.5)	-	0.167
HERO-2	3.0 (2.7)	2.8 (2.5)	2.9 (2.6)	3.4 (2.9)	2.5 (2.5)	1.9 (2.1)	<0.001
ASSENT-2	2.7 (2.5)	2.7 (2.5)	2.5 (2.5)	-	2.3 (2.5)	-	<0.001
ASSENT-3	2.6 (2.5)	2.6 (2.5)	3.3 (2.7)	-	2.5 (2.5)	-	0.014
All trials	2.8 (2.6)	2.8 (2.6)	2.9 (2.6)	3.4 (2.9)	2.5 (2.5)	1.9 (2.1)	<0.001
TIMI risk score				***Mean (SD)***			
GUSTO-1*	3.0 (2.2)	3.0 (2.2)	2.9 (2.2)	-		-	0.548
GUSTO-2b†	3.3 (2.2)	3.3 (2.2)	-	-		-	-
GUSTO-3	3.2 (2.2)	3.2 (2.2)	3.3 (2.2)	-	3.4 (2.2)	-	0.468
HERO-2	3.4 (2.3)	3.0 (2.2)	3.3 (2.2)	3.8 (2.4)	3.2 (2.2)	3.2 (1.9)	<0.001
ASSENT-2	2.8 (2.2)	2.8 (2.2)	2.7 (2.2)	-	2.9 (2.1)	-	0.273
ASSENT-3	2.8 (2.1)	2.7 (2.1)	3.5 (2.3)	-	3.0 (2.1)	-	<0.001
All trials	3.1 (2.2)	3.0 (2.2)	3.2 (2.2)	3.8 (2.4)	3.1 (2.2)	3.2 (1.9)	<0.001

*Excludes t-PA + streptokinase + IV heparin treatment arm. †ST-segment elevation stratum only. ‡*P*-value for chi-squared test of association between region and 30-day mortality. West = Western countries; EE = Eastern Europe; LA = Latin America. SD = Standard deviation.

6.3.3 Mortality by trial

Table 6.5 shows 30-day mortality rates by trial and region. Overall mortality in HERO-2 was higher due to high recruitment in non-Western countries. In the other trials mortality was also generally higher in the non-Western regions, more so in Latin America. In Western countries crude mortality rates were lowest in ASSENT-3 (5.7%) and highest in GUSTO-3 (7.3%), which correlates with average overall risk (Table 6.4).

Aside from differences in overall risk levels variations in mortality may also be attributable to differences in the type of fibrinolytic therapy used, as considered earlier. Also there were differences in eligibility and inclusion criteria; the MI's suffered in the ST-segment elevation stratum of GUSTO-2b may have been less severe and as a result there were fewer deaths.

The mortality rate for Eastern Europe in HERO-2 appears to be an outlier. Average risk profiles for these patients were no higher than GUSTO-3 and ASSENT-3 patients in this region. The HERO-2 mortality rate for this region should be considered as reliable as it is based on almost 6000 patients, which is substantially more than the other trials.

Table 6.5 Thirty-day mortality by trial and region

		Overall		Region					
				West	EE	Russia	LA	Asia	
Trial	**N**	***Deaths***	**%**	***Deaths* (%)**					***P*‡**
GUSTO-1*	30647	2128	6.9	6.9	8.8				0.22
GUSTO-2b†	4130	250	6.1	6.1					-
GUSTO-3	15059	1113	7.4	7.3	8.7		12.4		0.03
HERO-2	17073	1850	10.8	6.7	10.2	13.2	10.8	11.0	<0.001
ASSENT-2	16949	1045	6.2	6.0	5.9		10.4		<0.001
ASSENT-3	6095	364	6.0	5.7	6.7		8.9		0.01
All trials	89953	6750	7.5	6.6	9.7	13.2	10.4	11.0	<0.001

*Excludes t-PA + streptokinase + IV heparin treatment arm. †ST-segment elevation stratum only. ‡*P*-value for chi-squared test of association between region and 30-day mortality. West = Western countries; EE = Eastern Europe; LA = Latin America.

6.3.4 Comparison of models

Table 6.6 shows a comparison of the HERO-2, ASSENT-2 and GUSTO-I model coefficients. The ASSENT-3 dynamic model is shown in Appendix B Table A2. There is a notable similarity in the effect of coefficients across the trials. The HERO-2 model re-calculated with the **angina** term omitted is presented to show the effect on the remaining regression coefficients of not adjusting for **angina** (see Methods Section 6.2.3).

The derived ASSENT-2 model is very similar to the HERO-2 model but fewer variables were selected; **diabetes** and **time from onset to randomisation** were eliminated in the backward step-down procedure. Confirming the findings of the HERO-2 model, significant **region** effects were also found, with non-Western patients associated with increased mortality risk.

Compared to the HERO-2 model the effect of **SYSBP** was estimated as larger in GUSTO-I and ASSENT-2 (for **SYSBP** below 120 and 140 respectively). Conversely the effect of **heart rate** in GUSTO-I and ASSENT-2 was smaller, at least in the 70-100 range. For the GUSTO-I model the **age** by **Killip class** interaction involved less attenuation of the **age** effect in higher **Killip classes** than in HERO-2 and ASSENT-2.

The increased risk associated with the risk factors **prior MI** and **hypertension** was estimated as larger in GUSTO-I and ASSENT-2 respectively.

A smaller effect for female **gender** was inferred by the ASSENT-2 model compared to HERO-2.

The increased risk associated with increasing **time from symptom onset to randomisation** was estimated as larger in GUSTO-I compared to HERO-2 for times above 2 hours.

The GUSTO-I model encompasses several effects not included in the HERO-2 model. These include **smoking status**, **prior CABG**, **cerebrovascular disease**, **height** and **weight**. To gauge whether or not

Table 6.6 Models as developed within each separate trial

Model:	Variable:					
		Region:				
	Intercept	Western countries	Latin America	Eastern Europe	Russia	Asia
HPI-FULL	-8.099	0	0.562	0.320	0.279	1.012
-minus angina	-8.089	0	0.576	0.351	0.345	1.004
ASSENT-2	-6.364	0	0.750	0.091	-	-
GUSTO-1	1.844	0	-	-	-	-

	Killip class:				**Slope of age (yrs):**			
	I	II	III	IV	I	II	III	IV
HPI-FULL	0	2.450	5.071	6.414	0.077	0.050	0.019	0.010
-minus angina	0	2.452	5.056	6.379	0.078	0.051	0.020	0.011
ASSENT-2	0	3.150	5.102	5.947	0.082	0.045	0.023	0.023
GUSTO-1	0	2.080	3.623	4.039	0.076	0.055	0.041	0.044

	Systolic BP (mm Hg)	**Heart rate (bpm)**	**Time to rx (hrs)**	**Female**
HPI-FULL	-0.018	0.030 (70-100)*	0.038	0.411
-minus angina	-0.018	0.030 (70-100)*	0.038	0.412
ASSENT-2	-0.027(≤140)*	0.018	-	0.225
GUSTO-1	-0.040(≤120)*	0.018 (50+)†	0.093 (2+)*	-

	MI location:			**Smoking status:**		
	Anterior	Inferior	Other	Never	Ex-smoker	Current
HPI-FULL	0	-0.509	-0.306	-	-	-
-minus angina	0	-0.512	-0.312	-	-	-
ASSENT-2	0	-0.399	-0.255	-	-	-
GUSTO-1	0	-0.536	-0.260	0	-0.213	-0.219

	History:					
	Angina	**CABG**	**CVD**	**MI**	**Diabetes**	**HTN**
HPI-FULL	0.229	-	-	0.249	0.270	0.130
-minus angina	-	-	-	0.320	0.276	0.164
ASSENT-2	-	-	-	0.227	-	0.239
GUSTO-1	-	0.352	0.341	0.412	0.250	0.165

	Fibrinolytic therapy:			
	Accelerated t-PA	SK+IV heparin	SK+SC heparin	t-PA+SK+IV heparin
GUSTO-1	-	0.214	0.197	0.140

	Weight (kg)	**Height (cm)**
GUSTO-1	-0.0074	$-0.03972HT+0.0001835(HT-154.9)_+^3 -0.0008975(HT-165.1)_+^3$ $+0.001587(HT-172.0)_+^3 -0.001068(HT-177.3)_+^3 +0.0001943(HT-185.4)_+^3$

*Values outside range truncated. †Linear spline with 1 knot at 50 (subtracted 1.968 from GUSTO-I model intercept). CVD = cerebrovascular disease; SK = streptokinase; SC = subcutaneous; rx = randomisation; HT = height; HTN = hypertension.

these variables were dropped from the HERO-2 model due to lesser power or just weaker associations we refitted the HERO-2 model including these extra factors. The estimated effect for **prior cerebrovascular disease** was almost as large (OR = 1.41 vs. 1.37). For **smoking status** there was no independent association, although when **gender** is omitted from the model this predictor becomes significant (*P*=0.01) and further the coefficients are very similar to the GUSTO-I model. This suggests that **smoking status** was included in the GUSTO-I model because **gender** was not adjusted for. For the other variables relations with mortality were much weaker.

In view of the apparent differences in some coefficient effects we compared the relative importance of predictors between models. To do this we ranked predictors according to how much prognostic information they independently contributed to each model. These results are shown in Table 6.7. In general all models identified **age**, **SYSBP**, **Killip class**, **heart rate**, **location of MI** and the **age by Killip class** interaction as the most significant independent predictors of mortality. Discrepancies included **prior MI** and **time to treatment** which were more important in the GUSTO-I model, consistent with their larger effect estimates in this model. Also **hypertension** ranked more highly in the ASSENT-2 model, consistent with its larger effect size. Lastly **gender** was more important in the HERO-2 model than in ASSENT-2 and its effect size was double the magnitude.

It is well known that coefficients (and their strength) are affected by the other explanatory/risk variables included in the model. For example a variable may have a stronger coefficient in one model because it is a proxy for other correlated effects which were not adjusted for but were included in the comparison model. This means that models can only be compared in the context of the variables which were adjusted for. This may be an explanation as to why **prior MI** was a much stronger predictor in the GUSTO-I model, as **angina** was not adjusted for which is correlated with **prior MI**. Also **hypertension** had a much stronger effect in the ASSENT-2 model, which is also correlated with **angina** a risk factor not considered for this model.

The issue remaining:

Despite models differing in their assessment of risk factors and their importance, are predictors of risk, depending on the model adopted, substantially different among participants in different trials and regions. Further, can re-calibration to the target population make models better applicable?

Table 6.7 **Independent contribution of each variable to each model. Variables are ranked in order of importance in terms of how much they contribute after adjustment for all other factors in the list.**

GUSTO-1*			HERO-2			ASSENT-2		
Variable	Adjusted χ^2	Rank	Variable	Adjusted χ^2	Rank	Variable	Adjusted χ^2	Rank
Age, y†	717	1	Age, y†	541	1	Age, y†	501	1
SYSBP, mm Hg	550	2	SYSBP, mm Hg	218	2	SYSBP, mm Hg	180	2
Killip class†	350 (3 df)	3	Heart rate, bpm	151	3	Heart rate, bpm	110	3
Heart rate, bpm	275 (2 df)	4	Killip class†	101 (3 df)	4	Killip class†	55 (3 df)	4
Location of MI	143 (2 df)	5	Location of MI	78 (2 df)	5	Age-by-Killip-class	35 (3 df)	5
Prior MI	64	6	Age-by-Killip-class	57 (3 df)	6	Location of MI	31 (2 df)	6
Age-by-Killip-class	29 (3 df)	7	Gender	47	8	Region	25 (4 df)	7
Height, cm	31 (4 df)	8	Region	48 (4 df)	7	Hypertension	12	8
Time to treatment, h	23	9	Diabetes	15	9	Gender	9	9
Diabetes	21	10	Angina	14	10	Prior MI	7	10
Weight, kg	16	11	Prior MI	12	11			
Smoking	22 (2 df)	12	Hypertension	5	12			
Thrombolytic therapy	15 (3 df)	13	Time to treatment, h	3	13			
Previous CABG	16	14						
Hypertension	14	15						
Prior CVD	10	16						

*Taken from Lee et al [1].

†The likelihood ratio tests for age and Killip class were conducted following the same approach taken in the GUSTO-I model publication, i.e., the age-by-Killip-class interaction was not dropped from the reduced model. We also conducted the tests for these variables by removing both the main effect and interaction term and found that age and Killip class were still among the five most important predictors in both the HERO-2 and ASSENT-2 models.

CVD = cerebrovascular disease.

6.3.5 Performance of models

6.3.5.1 Discriminatory ability

Discriminatory ability results for "HPI-FULL – angina" were very similar to HPI-FULL and so are shown in Appendix B.

Table 6.8 shows the overall results for the discriminatory ability of models in each VIGOUR dataset. Discrimination was excellent as *c* statistics were around 0.81 for all models. There is no evidence of row or column effects; no model performed consistently better and performance was not better in any one trial or group of trials. At the cell level it can be seen that within a particular trial dataset (column) the model that was developed on that dataset generally performed best, although only a little better than other models. Comparing across trials (rows), models achieved the best results in the dataset they were developed on, however their performance was almost the same in other datasets. Differences in average performance by model were very slight.

Table 6.8 Overall *c* statistic results

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med‡**
HPI-FULL*	0.818	0.809	0.807	0.811	0.817	0.818	0.814
GUSTO-1†	0.812	0.808	0.803	0.819	0.809	0.825	0.811
ASSENT-2*	0.809	0.813	0.809	0.811	0.820	0.821	0.812
Median	0.812	0.809	0.807	0.811	0.817	0.821	0.812

*Treatment offset incorporated.

†GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

‡Median.

The results by region are shown in Appendix B Table A3. There is no consistent evidence of an advantage of a particular model or models in a particular region, hence no evidence of an interaction between model and region.

Since the performance of models within each trial-region subset was very consistent we have summarised these results by taking the median model *c* statistic (Table 6.9). There were particular trial effects; *c* statistics increased to around 0.86 in GUSTO-3 for Latin America and in ASSENT-3 for Eastern Europe. In HERO-2 all models suffered a decline in performance when applied to Asian patients; *c* statistics dropped to around 0.76. This may be evidence of the importance of region to performance independent of model within a particular trial dataset. However the increases observed in GUSTO-3 and ASSENT-3 may be consistent with chance as patient numbers were small.

On independent datasets from a later related trial the related models most often performed best, however only minor gains in discriminatory capacity were observed. For example, the ASSENT-2 and

Table 6.9 Median *c* statistics (3 models), by region

	Population dataset						
Region	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med***
Western	0.813	0.810	0.806	0.813	0.817	0.822	0.813
EE	0.797	0.788	0.863	0.744		0.805	0.797
Russia	0.822						0.822
LA	0.824	0.786	0.799			0.859	0.812
Asia	0.766						0.766

Western = Western countries; EE = Eastern Europe; LA = Latin America.

*Median.

GUSTO-I models performed best overall and in all regions except Latin America when applied in ASSENT-3 and GUSTO-3 respectively.

The results also demonstrate that models developed in primarily Western patients, i.e. the GUSTO-I and ASSENT-2 models, hold up well in Eastern Europe and Russia. Conversely, as demonstrated by the HERO-2 model, a model developed in predominantly Eastern Europe and Russia holds up well when applied in the West. Concerning the discriminatory ability of the ASSENT-3 dynamic model at day 0 (i.e., baseline version of the model), a *c* statistic of 0.82 was reported for apparent performance [9]. The discriminatory performance of the models here when applied in ASSENT-3 was only slightly less. Furthermore the performance of the HERO-2 and GUSTO-I models in their development datasets was very similar to the ASSENT-3 model.

The application of trial-region level recalibration allows patient pairs from different regions to be ranked more accurately. Accordingly slight improvements were observed mainly for the GUSTO-I model since this model includes no region effects (Table A4). The general pattern remained consistent with results for un-calibrated models.

6.3.5.2 R^2 comparisons

R^2 results for "HPI-FULL – angina" were very similar to HPI-FULL and so are shown in Appendix B.

Overall R^2 results before the application of recalibration are shown in Table 6.10. There is evidence of a column (trial) effect as the performance of models was similar within trials but there was variation between trials. In HERO-2 and GUSTO-3 increases in explained variability were noted across all models as reflected in the higher median R^2 values. There is no convincing evidence of a row (model) effect; the performance of models was similar on average.

At the cell level, looking within columns, again the model that was developed on or related to a particular trial achieved the best R^2. An exception was the poor performance of the GUSTO-I model in GUSTO-2b. The similar GUSTO-2b and ASSENT-2 trial mortality rates partially explain the

superiority of the ASSENT-2 model in GUSTO-2b. Comparisons within rows are not meaningful because of the apparent trial effect.

Table 6.10 Overall R^2 statistics (%) for un-calibrated models

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med‡**
HPI-FULL*	26.9	20.6	20.4	22.6	20.1	23.6	21.6
GUSTO-1†	23.6	19.5	19.7	24.0	14.6	24.0	21.7
ASSENT-2*	24.4	21.3	21.2	22.7	22.4	24.0	22.6
Median	24.4	20.6	20.4	22.7	20.1	24.0	21.7

*Treatment offsets incorporated in all models.

†GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

‡Median.

Table A5 in Appendix B shows the results within regions. There were some minor deviations from the overall results pattern concerning models performing better in unrelated trials; for example in HERO-2 the GUSTO-I model performed best in Western patients. However there was no consistent evidence for a model having an advantage in a particular region.

In general models performed similarly within trial-region subgroups so we have summarised results by taking medians (Table 6.11). One major exception to this was in Asia where the performance of the GUSTO-I and ASSENT-2 models was poor compared to the HERO-2 model. Within regions (row) there were rises and falls in average performance across trials, however there was a suggestion of enhanced performance in Latin America as the overall median was high. Average model performance was highest and lowest in Russia and Asia respectively, but these trends are unsubstantiated as only one trial recruited patients in these regions.

Table 6.11 Median (3 models) R^2 statistics (%) for un-calibrated models, by region

	Population dataset						
Region	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Median**
Western	21.2	20.2	20.0	22.8	20.1	24.0	20.7
EE	21.5	20.3	29.3	14.4		21.4	21.4
Russia	27.3						27.3
LA	28.4	23.0	20.3			34.0	25.7
Asia	10.8						10.8

Western = Western countries; EE = Eastern Europe; LA = Latin America.

Explained variability in independent datasets improved after trial-level recalibration and within trial differences between models reduced (Table 6.12). Results by region are shown in Appendix B Table

A6. Due to the recalibration performance declined for some models in particular trial-region subsets. In HERO-2 Western patients the GUSTO-I and ASSENT-2 model R^2's decreased by 1.9 and 1.2% respectively. HERO-2 model performance decreased slightly (<0.5%) in Eastern Europe across all trials that recruited in this region. Large differences remained in Asia with the ASSENT-2 model still performing substantially worse.

Table 6.12 Overall R^2 statistics (%) after trial-level recalibration

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med‡**
HPI-FULL*	26.9	20.7	20.4	22.7	21.8	23.7	22.3
GUSTO-1†	25.1	19.8	19.9	24.0	18.6	24.3	22.0
ASSENT-2*	25.3	21.3	21.2	22.7	23.2	24.2	23.0
Median	25.3	20.7	20.4	22.7	21.8	24.2	22.3

*Treatment offsets incorporated in all models before recalibration.

†GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

‡Median.

Table 6.13 shows median model results for each trial-region subset. The pattern is very similar to that observed prior to recalibration.

Table 6.13 Median (3 models) R^2 statistics (%) obtained after trial-level recalibration, by region

	Population dataset						
Region	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Median**
Western	20.6	20.3	20.1	22.8	21.8	24.2	21.2
EE	22.5	20.5	28.8	14.2		21.7	21.7
Russia	28.9						28.9
LA	26.9	23.5	21.2			34.0	25.2
Asia	17.1						17.1

Western = Western countries; EE = Eastern Europe; LA = Latin America.

Table 6.14 shows overall results after the application of trial-region calibration. Within trial differences diminished further. Notwithstanding, the remarks made earlier for uncalibrated models still hold: e.g., within trials the model that was developed on or is related to a trial performed best.

Table A7 in Appendix B shows results by region. Again differences between models within trial-region subgroups decreased. Despite the regional adjustment which places all models on an even playing field, usually the model that was developed from or was related to that trial-region subset still performed best. So no model displayed a consistent advantage in a particular region. Interestingly the

Table 6.14 Overall R^2 statistics (%) after trial-region level recalibration

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med‡**
HPI-FULL*	26.9	20.7	20.4	22.7	21.8	23.7	22.3
GUSTO-1†	25.5	20.1	20.2	24.1	18.6	24.4	22.2
ASSENT-2*	26.1	21.3	21.3	22.8	23.2	24.2	23.0
Median	26.1	20.7	20.4	22.8	21.8	24.2	22.3

*Treatment offsets incorporated in all models before recalibration.

†GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

‡Median.

GUSTO-I model performed best in Latin America of both ASSENT trials and in Asia; it was not derived from these regions. Perhaps the additional variables it includes afford it this extra explanatory ability.

There were a few instances where explained variability was slightly lower than for un-calibrated models. This relates to the original coefficient effects being more optimal than the recalibrated coefficients in certain regions. The updated coefficients are the product of the calibration slope and the original coefficient values [57]. For example the GUSTO-I model coefficients are based on Western patients. Recalibration in the HERO-2 trial means these coefficients are updated using a global sample. In HERO-2 Western patients this updated model did not perform quite as good as the un-calibrated model.

In Asia the performance of all models was now consistent.

Table 6.15 shows median model results for each trial-region subset. Average performance increased across the board. The overall pattern remains consistent with that observed prior to recalibration.

Table 6.15 Median (3 models) R^2 statistics (%) obtained after trial-region level recalibration, by region

	Population dataset						
Region	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Median**
Western	21.2	20.3	20.1	22.8	21.8	24.2	21.5
EE	22.5	20.5	30.3	14.2		22.0	22.0
Russia	29.1						29.1
LA	28.7	24.4	22.5			34.6	26.6
Asia	18.8						18.8

Western = Western countries; EE = Eastern Europe; LA = Latin America.

The magnitude of the R^2 statistics obtained here are similar to those reported by Steyerberg et al [66] who, using the GUSTO-I dataset, studied interval validation procedures for logistic regression

predicting 30-day mortality following AMI. By conducting repeated sampling the apparent and test performance of an eight-predictor model was assessed in data sets with 5 to 80 events per variable. Average R^2 values of around 20% were obtained for datasets with EPV's of at least 20. The R^2 for this model when fitted to the total GUSTO-I dataset was also 20%.

6.3.5.3 Intercept and calibration slope assessment

Overall results for model performance using the intercept and slope assessment are shown in Table 6.16 for un-calibrated models.

The HERO-2 model HPI-FULL calibrated well in all trial datasets except GUSTO-2b, however all models significantly miscalibrated in this dataset. Downward adjustment of the average risk predicted by both HERO-2 models would be sufficient to rectify this, as calibration slope estimates were close to one.

Omission of the risk factor **angina** from the HERO-2 model caused significant underestimation of risk in GUSTO-3. In GUSTO-I the overall test of miscalibration for this model was not significant, although the calibration slope was significantly higher than one (i.e., OR = 1.04; 95% CI, 1.00-1.09). This indicates that inflation of the expected relationship between the covariables and risk would improve predictions. The effect of inflation is to make predictions more extreme as coefficients are up weighted.

Table 6.16 Intercept (α) and calibration slope (β) for un-calibrated models

	Population dataset								
	HERO-2			ASSENT-2			ASSENT-3		
Model	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	-0.005	1.00	0.98	0.055	0.99	0.06	0.014	0.99	0.73
HPI-FULL*†	-0.008	1.00	0.95	-0.064	0.98	0.68	-0.105	0.97	0.61
GUSTO-1‡	0.139	0.91§	<0.001	-0.221	0.87§	<0.001	-0.141	0.91	0.06
ASSENT-2*	0.184	0.97	<0.001	0.000	1.00	0.99	-0.025	1.01	0.70
	GUSTO-I			**GUSTO-2**			**GUSTO-3**		
Model	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	0.078	1.04	0.15	-0.368	0.99	<0.001	0.122	1.03	0.08
HPI-FULL*†	0.092	1.04§	0.13	-0.324	0.99	<0.001	0.145	1.03	0.03
GUSTO-1‡	-0.037	0.98	0.49	-0.862	0.76§	<0.001	0.042	0.96	<0.001
ASSENT-2*	0.056	1.04	0.13	-0.066	1.08	0.002	0.216	1.05	0.002

*Treatment offsets incorporated in all models.

†HPI-FULL re-estimated excluding angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

§Calibration slope estimate significantly different from 1 indicating either shrinkage or inflation is required.

The GUSTO-I model miscalibrated significantly in all independent data except in ASSENT-3, although the corresponding p-value was borderline significant. In most cases shrinkage of regression coefficients was indicated to correct this miscalibration, even in GUSTO-2b a related trial. The ASSENT-2 model performed somewhat better as it calibrated adequately in ASSENT-3 and GUSTO-I, but was not as good as HPI-FULL. HPI-FULL was the only model to pass the test of miscalibration in GUSTO-3.

The significant over-estimation of risk by all models in GUSTO-2b appears to be due to patients having suffered less severe MI's which stems from this trial having a less stringent eligibility criteria.

Table 6.17 shows this assessment for trial-region subgroups which demonstrates that in some cases miscalibration is due to regional effects. For example in HERO-2 the GUSTO-I and ASSENT-2 model predictions were highly inaccurate for the whole sample, as may be expected based on the trial population, but performed very well in Western patients. The accuracy of the ASSENT-2 model was also satisfactory in Latin America.

The accuracy of HPI-FULL predictions was preserved in all regions with the possible exception of Western countries in GUSTO-3; the p-value reached borderline significance (P=0.05). The same was true for HPI-FULL with **angina** excluded.

Calibration slopes were very similar for the HERO-2 and ASSENT-2 models across the board indicating these models project a similar relationship between patient risk factors and mortality. In GUSTO-I Western patients the slope estimates for the HERO-2 models indicated inflation was needed. The corresponding ASSENT-2 model slope was not significant, but was of similar magnitude.

For the GUSTO-I model shrinkage of coefficients was indicated in Eastern Europe and Russia in HERO-2 as slope estimates were significantly less than one. Similar slope estimates were obtained in the other trials for this region, excluding ASSENT-3 which recruited a small number in Eastern Europe. Shrinkage factors of around 0.88 were obtained meaning that regression coefficients need to be shrunk by 12% to obtain better calibrated predictions. The same amount of shrinkage is indicated for this model in Western patients of both ASSENT trials.

The slope results suggest the association between patient risk factors and mortality is stronger in GUSTO-I.

The application of trial-level recalibration was not satisfactory as problems still occurred within some regions as shown in Table 6.18. In HERO-2 the upward adjustment of risk predicted by the GUSTO-I and ASSENT-2 models resulted in significant over-estimation in Western patients. Risk prediction was also inaccurate in the other HERO-2 regions; the GUSTO-I model failed in Asian patients and worse still the ASSENT-2 model calibrated poorly in all regions except Eastern Europe. The GUSTO-I model also underestimated risk in Latin American patients of both ASSENT trials. In all cases slope estimates were not significantly different from one. This means the application of trial-level recalibration requires region-specific intercepts but region-specific slopes are not warranted.

Table 6.17 Intercept (α) and calibration slope (β)[§] results for un-calibrated models within regions

Population:	Western countries:																	
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-I			GUSTO-2			GUSTO-3		
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	0.119	1.06	0.76	0.052	0.99	0.10	0.077	1.01	0.59	0.085	**1.05**	0.11	-0.368	0.99	<0.001	0.145	1.03	0.05
HPI-FULL*†	0.111	1.06	0.78	-0.064	0.98	0.71	-0.044	0.99	0.91	0.098	**1.05**	0.10	-0.324	0.99	<0.001	0.168	1.04	0.02
GUSTO-1‡	0.111	1.05	0.81	-0.260	**0.87**	<0.001	-0.209	**0.90**	0.13	-0.035	0.98	0.61	-0.862	**0.76**	<0.001	0.022	0.96	0.003
ASSENT-2*	0.125	1.07	0.70	0.006	1.00	0.99	0.055	1.03	0.85	0.061	1.04	0.08	-0.066	1.08	0.002	0.224	1.06	0.005
Population:	**Eastern Europe:**															**Russia:**		
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-I			GUSTO-3			HERO-2		
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	-0.045	0.98	0.86	-0.159	0.99	0.78	0.331	1.37	0.56	-0.402	0.77	0.45	-0.188	0.92	0.73	-0.021	0.99	0.95
HPI-FULL*†	-0.050	0.98	0.83	-0.314	0.98	0.41	0.101	1.34	0.44	-0.347	0.78	0.45	-0.143	0.94	0.83	-0.020	0.99	0.96
GUSTO-1‡	0.077	**0.87**	<0.001	-0.118	0.91	0.77	-0.123	1.01	0.96	0.005	0.79	0.06	0.180	0.91	0.01	0.115	**0.89**	<0.001
ASSENT-2*	0.169	0.95	<0.001	0.094	1.04	0.97	0.649	1.42	0.63	-0.241	0.74	0.14	0.119	0.94	0.15	0.333	0.98	<0.001
Population:	**Latin America:**												**Asia:**					
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-3			HERO-2					
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**			
HPI-FULL*	0.166	1.10	0.39	0.148	0.95	0.17	-0.266	0.88	0.65	0.498	1.23	0.65	-0.161	0.91	0.74			
HPI-FULL*†	0.157	1.10	0.41	0.028	0.94	0.48	-0.365	0.87	0.44	0.437	1.18	0.72	-0.172	0.90	0.71			
GUSTO-1‡	0.466	0.97	<0.001	0.475	0.90	<0.001	0.435	0.96	0.009	1.192	1.22	0.06	0.670	0.94	<0.001			
ASSENT-2*	-0.063	1.07	0.07	-0.089	0.95	0.90	-0.337	0.95	0.27	0.327	1.25	0.70	0.794	0.90	<0.001			

*Treatment offsets incorporated in all models.

†HPI-FULL re-estimated excluding angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

§Calibration slopes significantly different from 1 are shown in bold.

Table 6.18 Intercept (α) and calibration slope (β) results for recalibrated (trial-level) models, within regions

Population:	Western countries:																	
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-I			GUSTO-2§			GUSTO-3		
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	0.124	1.06	0.77	-0.003	1.00	0.99	0.062	1.02	0.90	0.006	1.00	0.99	0.000	1.00	1.000	0.022	1.01	0.95
HPI-FULL*†	0.119	1.06	0.78	-0.001	1.00	0.99	0.063	1.02	0.90	0.006	1.00	0.99	0.000	1.00	1.000	0.022	1.01	0.95
GUSTO-1‡	-0.051	1.16	<0.001	-0.040	1.00	0.58	-0.069	0.99	0.70	0.002	1.00	0.97	0.000	1.00	1.000	-0.020	1.00	0.80
ASSENT-2*	-0.079	1.11	0.001	0.006	1.00	0.99	0.080	1.02	0.80	0.005	1.00	0.97	0.000	1.00	1.000	0.007	1.01	0.95
Population:	**Eastern Europe:**															**Russia:**		
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-I			GUSTO-3			HERO-2		
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**
HPI-FULL*	-0.040	0.98	0.86	-0.214	1.00	0.54	0.311	1.39	0.51	-0.459	0.73	0.34	-0.297	0.90	0.42	-0.015	0.99	0.96
HPI-FULL*†	-0.042	0.98	0.85	-0.251	1.00	0.46	0.245	1.37	0.48	-0.415	0.74	0.34	-0.275	0.91	0.46	-0.011	0.99	0.98
GUSTO-1‡	-0.057	0.96	0.71	0.112	1.04	0.97	0.034	1.11	0.88	0.035	0.80	0.07	0.141	0.95	0.15	-0.022	0.98	0.90
ASSENT-2*	-0.012	0.99	0.90	0.094	1.04	0.97	0.685	1.41	0.66	-0.281	0.72	0.09	-0.073	0.89	0.32	0.147	1.02	0.02
Population:	**Latin America:**												**Asia:**					
	HERO-2			ASSENT-2			ASSENT-3			GUSTO-3			HERO-2					
Model:	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**	**α =**	**β =**	**P**			
HPI-FULL*	0.172	1.10	0.39	0.095	0.96	0.41	-0.279	0.89	0.65	0.351	1.20	0.76	-0.156	0.91	0.74			
HPI-FULL*†	0.166	1.10	0.42	0.089	0.96	0.45	-0.270	0.90	0.67	0.270	1.15	0.84	-0.164	0.90	0.71			
GUSTO-1‡	0.317	1.07	0.06	0.701	1.02	<0.001	0.585	1.06	0.03	1.139	1.27	0.12	0.525	1.03	0.002			
ASSENT-2*	-0.267	1.11	<0.001	-0.089	0.95	0.90	-0.313	0.94	0.38	0.072	1.18	0.71	0.622	0.93	<0.001			

*Treatment offsets incorporated in all models.

†HPI-FULL re-estimated excluding angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

§All patients in this trial were from the Western region so by definition estimates of intercept and calibration slope are perfect after re-calibration is applied.

By definition the application of trial-region level recalibration corrected the remaining bias in risk estimates due to regional differences (data not shown); therefore average expected risk matches observed risk in all regions for all models. Calibration slope estimates were roughly equivalent to that obtained for trial-level recalibration, as expected.

6.3.5.4 Calibration plots

The following plots depict observed versus predicted mortality according to deciles of predicted risk for each model, uncalibrated, and applied to each trial. The solid diagonal line reflects perfect calibration of the model predictions; the dashed line is the fitted smooth nonparametric lowess calibration curve.

The results are consistent with the analysis of intercept and calibration slope.

In HERO-2 both HERO-2 models over-estimated mortality in the lowest risk decile. Thorough scrutiny of this lack of agreement revealed zero deaths among Western patients in this risk decile for both models. Therefore the over-estimation was in Western patients but predictions were on average accurate for non-Western patients. The ASSENT-2 model predictions are under-estimated by a consistent amount (except in extreme low-risk patients) suggesting intercept adjustment is sufficient to rectify the bias. This accords with the intercept and slope analysis as the slope estimate was not significantly different from 1. For the GUSTO-I model the bias in risk probabilities was not as uniform across risk deciles.

In ASSENT-2 "HPI-FULL – angina" performed better than HPI-FULL as the fitted calibration curve followed the line of perfect calibration more closely. This is in agreement with the better miscalibration test result for "HPI-FULL – angina" (P=0.68 vs. 0.06). The GUSTO-I model generally under-estimated risk but the bias was worse in the lowest risk decile. This concurs with the slope estimate of 0.87 indicating that shrinkage is needed to make predictions less extreme.

In ASSENT-3 a very similar pattern was displayed by all models. Calibration was good except at the low-risk end of the distribution: in the lowest quintile mortality was under-estimated; in the third lowest decile mortality was over-estimated. The GUSTO-I model performed worst as also indicated by the test of miscalibration (P=0.06).

In GUSTO-I a similar pattern was displayed by all models; risk was under-estimated in the lowest risk decile but performance otherwise was very good.

In GUSTO-2b all models over-estimated risk. This over-estimation appeared more uniform for the HERO-2 models, an observation supported by slope estimates of approximately one. For the GUSTO-I model the observed bias was less for the lowest risk patients. Incidentally for this model the intercept and slope estimates were substantially less than 0 and 1 respectively. Therefore recalibration will affect the higher risk patients more as the effect of the shrinkage on the lowest risk LP scores will be largely offset by the large negative intercept, which concords with the plot.

In GUSTO-3 all of the models under-estimated risk apart from in the middle area of the distribution. The HERO-2 model displayed the best calibration which is confirmed by the only non-significant result for the test of miscalibration.

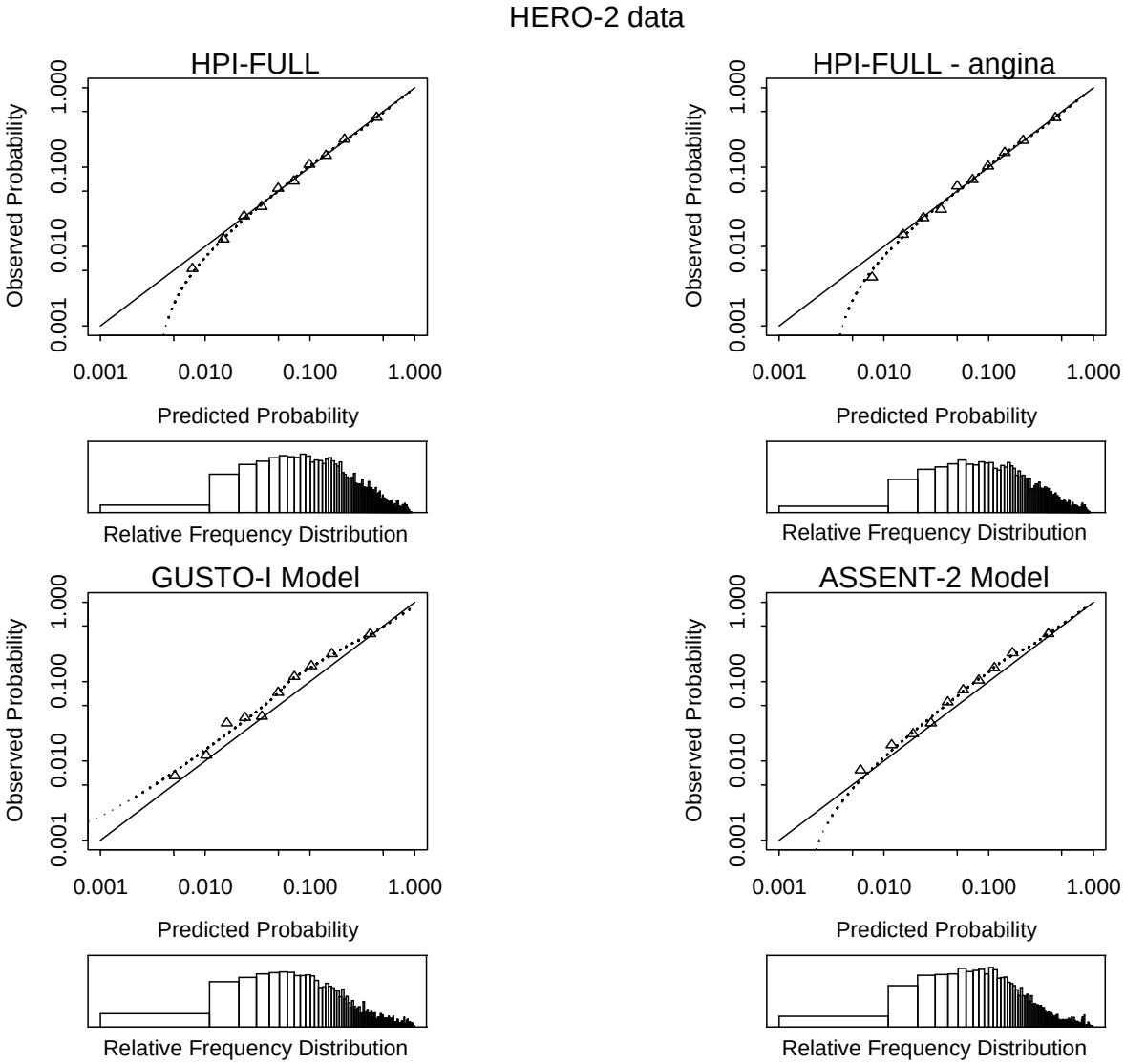

Figure 6.1 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in HERO-2 data.

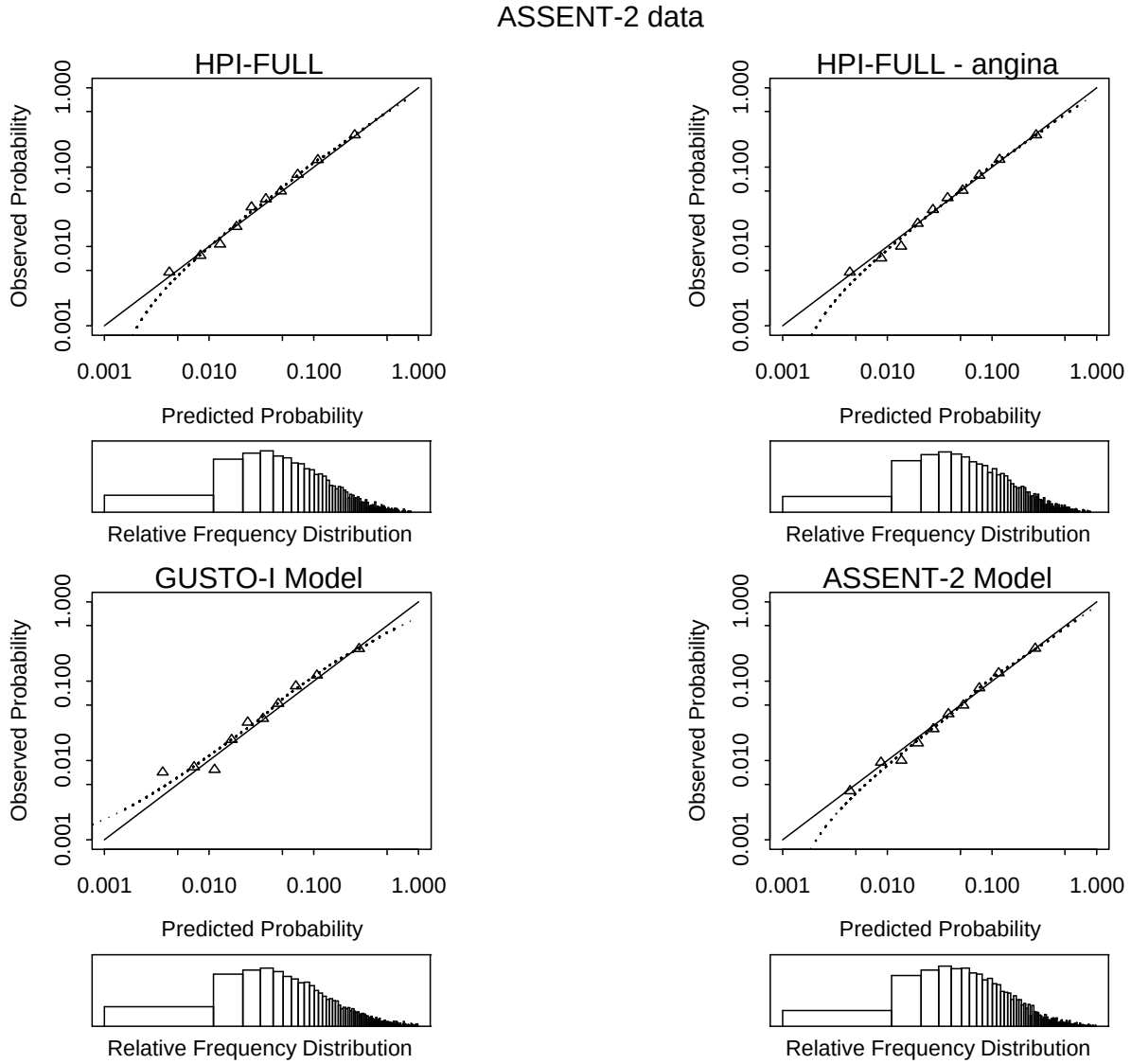

Figure 6.2 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in ASSENT-2 data.

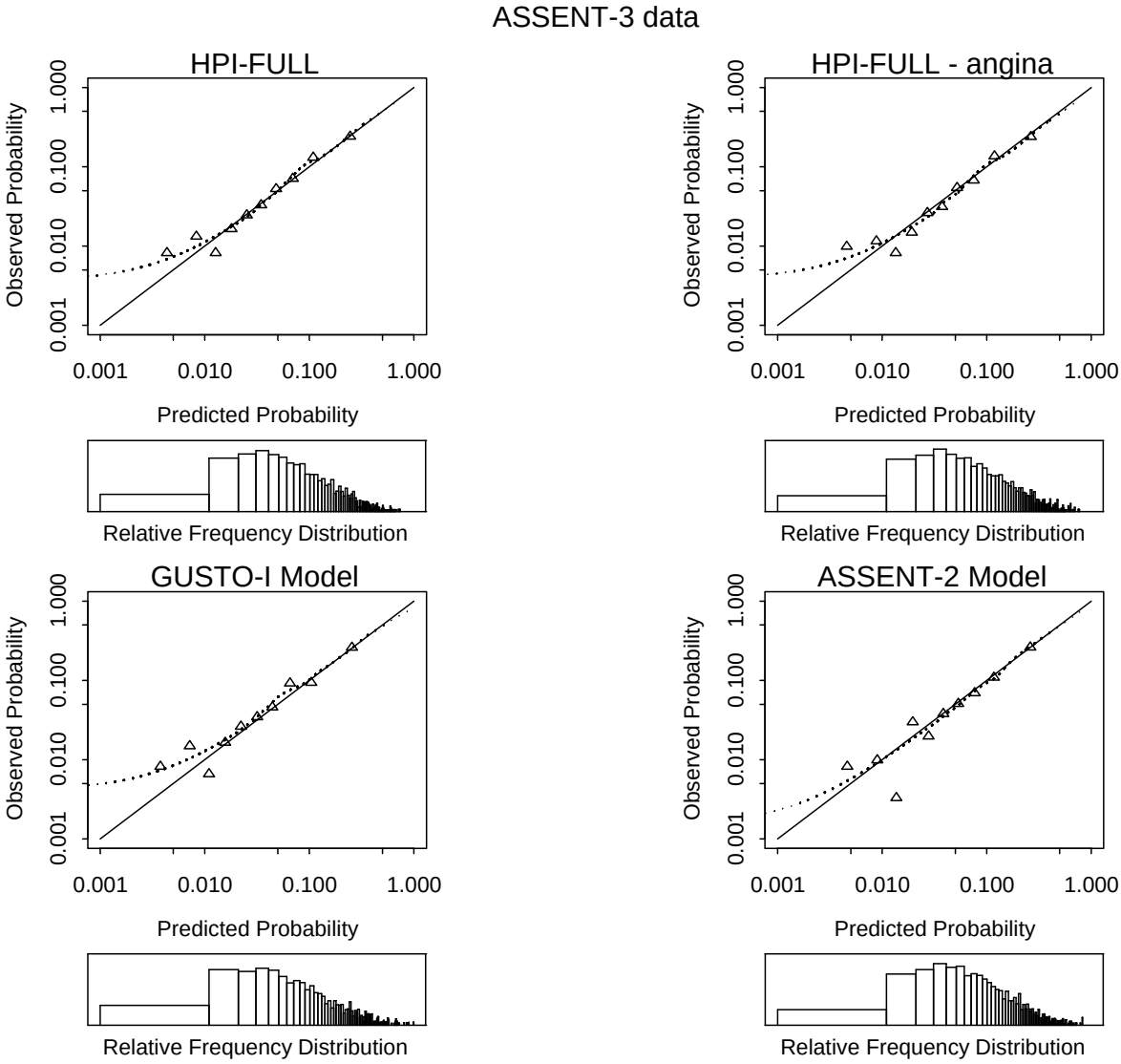

Figure 6.3 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in ASSENT-3 data.

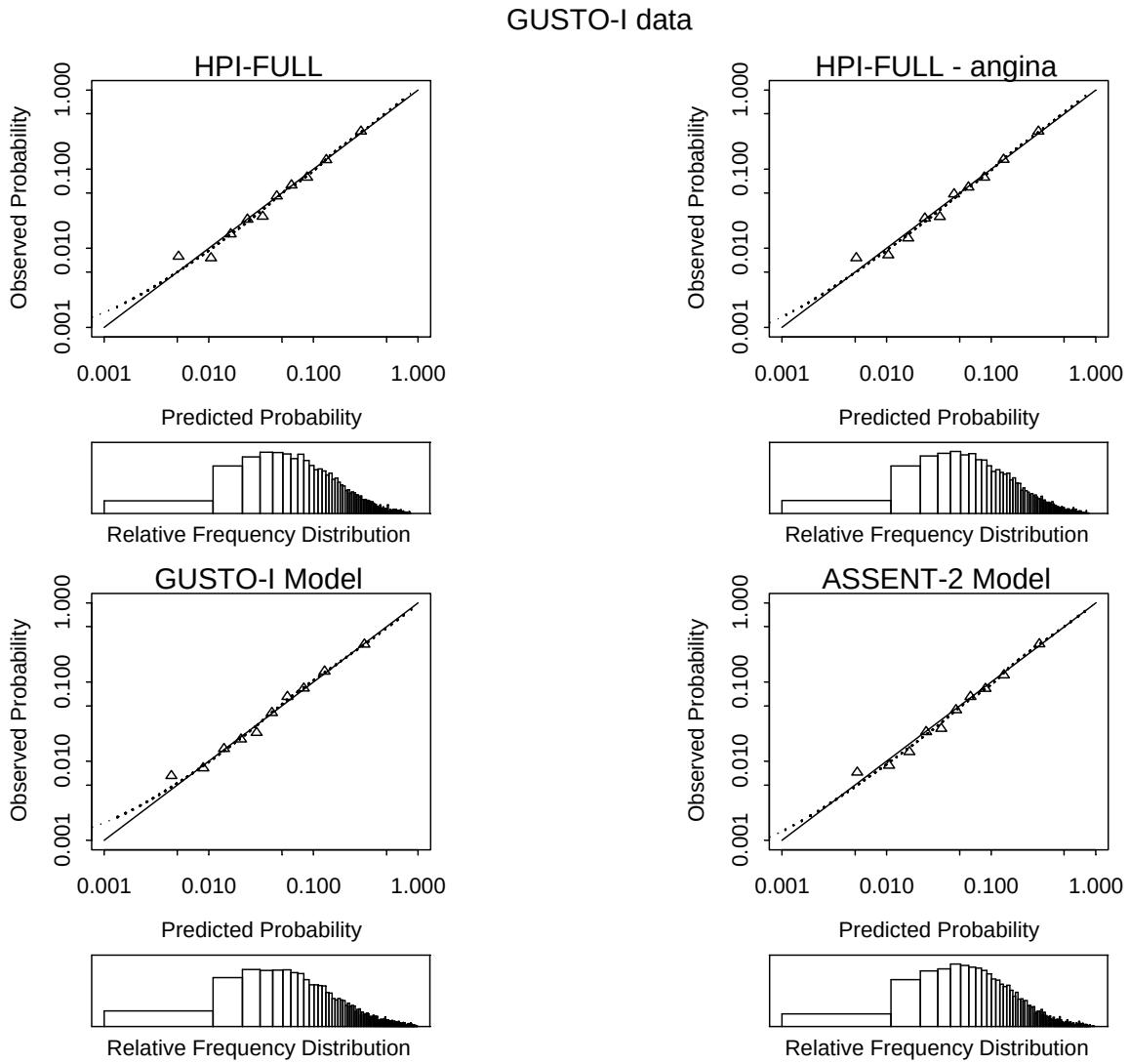

Figure 6.4 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in GUSTO-1 data.

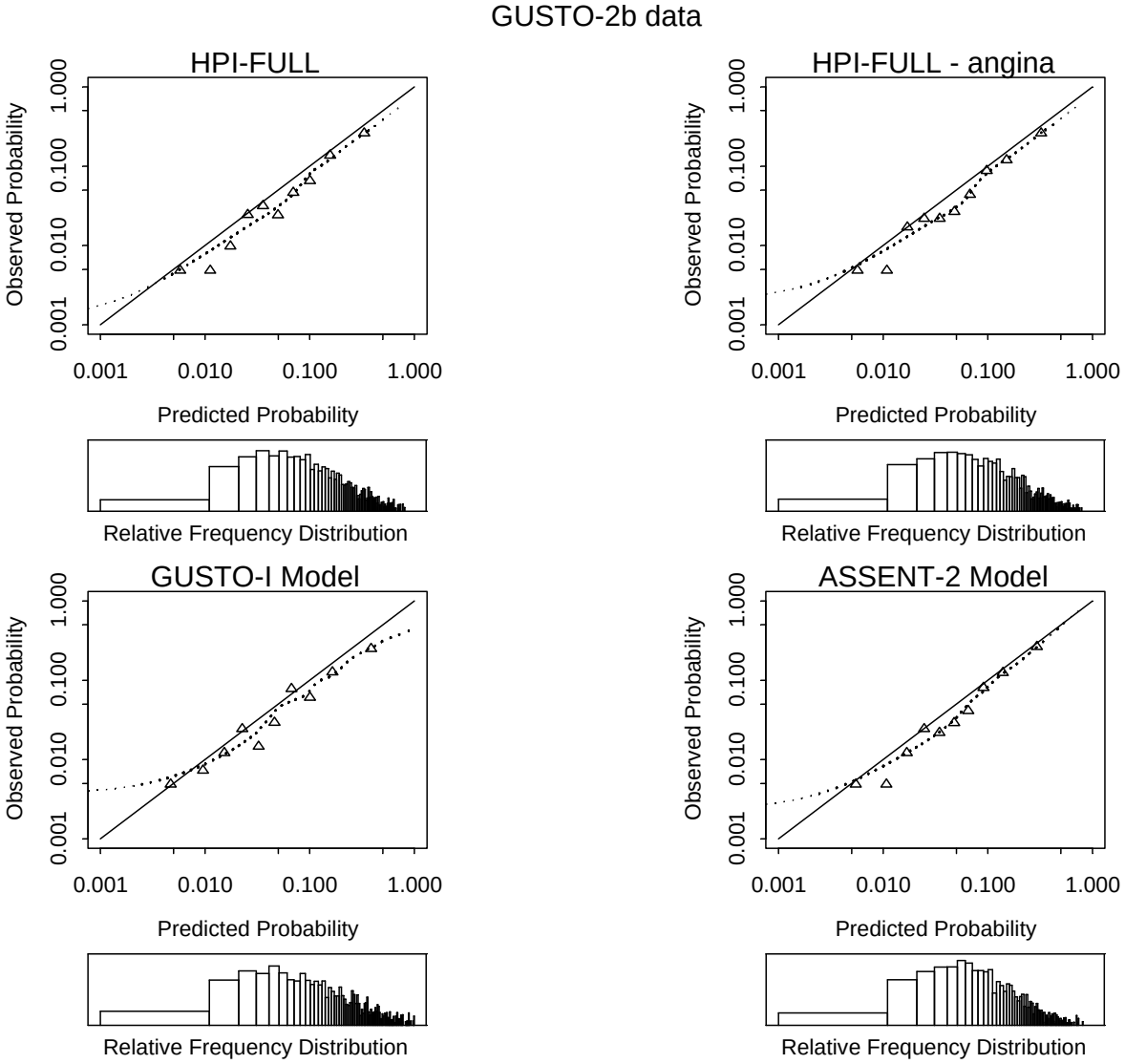

Figure 6.5 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in GUSTO-2b data.

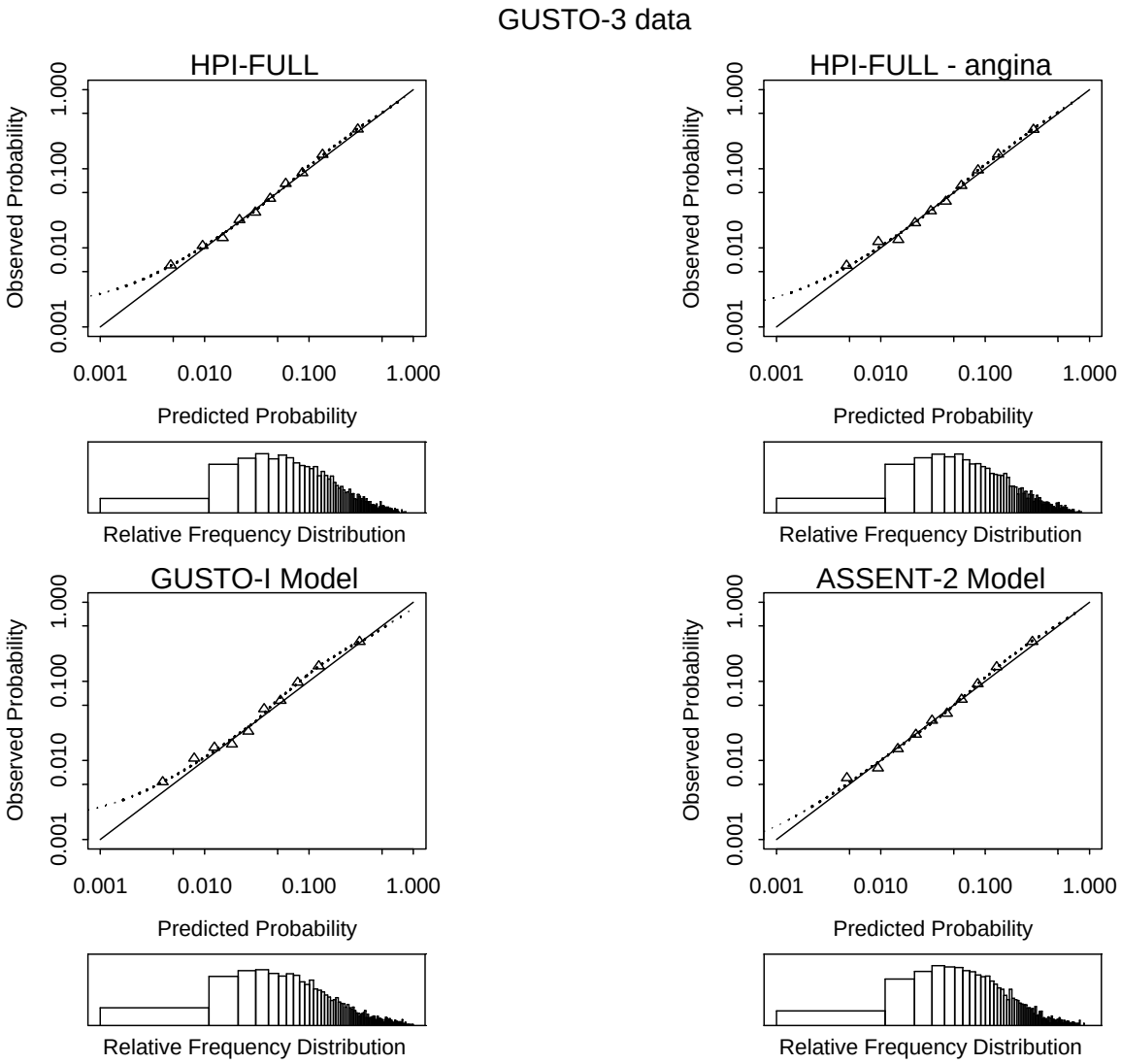

Figure 6.6 HERO-2, GUSTO-I and ASSENT-2 models applied as they are (i.e., un-calibrated) in GUSTO-3 data.

6.3.6 Performance of prognostic indices

6.3.6.1 Discriminatory ability

Table 6.19 shows *c* statistics for HPI-FULL and the TIMI and HPI risk scores, calibrated by trial and region. There is no evidence of a column effect, i.e., average performance is no better in any one trial or trials. However there is evidence of a row effect; the performance of the risk scores was lower compared to HPI-FULL. The difference in performance between the risk scores themselves was less. No risk strategy demonstrated a strong advantage in a particular trial; the relative difference between them was generally consistent across trials.

The loss in discriminatory ability observed in HERO-2 from using the HPI risk score in place of HPI-FULL was no worse in the other trials. The same degree of benefit observed in HERO-2 from using HPI-CAT in place of TIMI was not observed in all the other trials, although HPI-CAT was never any worse than TIMI.

Table 6.19 Overall *c* statistic results for HERO-2 full model, HPI and TIMI risk scores*

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med§**
HPI-FULL†	0.818	0.809	0.807	0.811	0.817	0.818	0.814
HPI-CAT‡	0.792	0.792	0.784	0.791	0.809	0.801	0.792
TIMI‡	0.781	0.791	0.784	0.783	0.794	0.798	0.788

Significant differences in *c* statistics: HPI-FULL vs. HPI-CAT $P<0.001$ in all trials except GUSTO-2b ($P=0.39$); HPI-CAT vs. TIMI $P<0.001$ in HERO-2 and GUSTO-I, $P>0.05$ in remaining trials.

*All risk strategies calibrated by trial and region.

†Treatment offset incorporated prior to re-calibration.

‡Treatment offset only incorporated for application in GUSTO-I.

§Median.

The results by region are shown in Table 6.20. Because the majority of patients were recruited in Western countries, excluding HERO-2, results for this region were very similar to the overall results. As anticipated the HERO-2 strategies displayed a noticeable advantage in HERO-2 regions. In both ASSENT trials there was very little difference in performance between all approaches in Latin America. In GUSTO-3 the increased discrimination noted earlier for models in Latin America (Table 6.9) was also apparent with the risk scores.

There is a suggestion that HPI-CAT is superior to TIMI in Eastern Europe, although in independent data patient numbers were small.

The loss in discrimination from using a risk score instead of a risk model is no worse in any particular region.

Table 6.20 *c* statistic results for HERO-2 full model, HPI and TIMI risk scores, by region*

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-2	GUSTO-3
Western countries:						
HPI-FULL†	0.813	0.810	0.806	0.812	0.817	0.818
HPI-CAT‡	0.800	0.791	0.780	0.792	0.809	0.799
TIMI‡	0.787	0.791	0.782	0.783	0.794	0.796
Latin America:						
HPI-FULL†	0.828	0.783	0.785			0.874
HPI-CAT‡	0.804	0.784	0.783			0.881
TIMI‡	0.794	0.784	0.787			0.847
Eastern Europe:						
HPI-FULL†	0.803	0.788	0.863	0.744		0.801
HPI-CAT‡	0.772	0.790	0.849	0.735		0.798
TIMI‡	0.761	0.775	0.778	0.711		0.814
Russia:						
HPI-FULL†	0.825					
HPI-CAT‡	0.795					
TIMI‡	0.780					
Asia:						
HPI-FULL†	0.766					
HPI-CAT‡	0.743					
TIMI‡	0.761					

*All models calibrated by trial and region.

†Treatment offset incorporated prior to re-calibration.

‡Treatment offset only incorporated for application in GUSTO-I.

6.3.6.2 R^2 results

Table 6.21 shows overall R^2 results for HPI-FULL and the TIMI and HPI risk scores, calibrated by trial and region. There is evidence of a row (i.e., model) effect; HPI-FULL performed better than the risk scores and HPI-CAT out-performed TIMI except in ASSENT-3. As expected the HERO-2 strategies performed best when applied in HERO-2 and the advantage from using HPI-CAT over TIMI was generally less outside of HERO-2. A small column effect is also shown as minor fluctuations in average explained variability were noted across the trials; results were lower in ASSENT.

Table 6.22 shows results by region. There were some deviations from the overall pattern but nothing consistent to conclude heterogeneity in performance according to region.

In HERO-2 Western countries R^2 values decreased from the overall results in a uniform fashion for all prediction methods. As expected, results changed little in the other trials.

In Latin America results were more heterogeneous: HPI-CAT outperformed HPI-FULL in GUSTO-3 and in ASSENT-3 both risk scores performed better than HPI-FULL.

In Eastern Europe all approaches suffered a decline in performance in GUSTO-I. In ASSENT-3 performance increased considerably for the HERO-2 strategies but stayed roughly consistent for TIMI. In GUSTO-3 both risk scores performed better than HPI-FULL with TIMI doing best.

Table 6.21 Overall R^2 statistics (%) for HERO-2 full model, HPI and TIMI risk scores*

	Population dataset						
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**	**Med§**
HPI-FULL†	26.9	20.7	20.4	22.7	21.8	23.7	22.3
HPI-CAT‡	22.4	18.5	17.5	19.6	21.4	21.1	20.4
TIMI‡	20.1	17.3	17.7	18.0	18.7	20.2	18.4
Median	22.4	18.5	17.7	19.6	21.4	21.1	20.4

*All models calibrated by trial and region.

†Treatment offset incorporated prior to re-calibration.

‡Treatment offset only incorporated for application in GUSTO-I.

§Median.

Table 6.22 Nagelkerke's R^2 statistics (%) for the HERO-2 full model, HPI and TIMI risk scores*

Model/Dataset	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**
Western patients:						
HPI-FULL†	21.2	20.3	20.1	22.8	21.8	23.6
HPI-CAT‡	19.3	17.9	16.6	19.6	21.4	20.8
TIMI‡	17.0	16.7	17.0	18.1	18.7	19.8
Latin America:						
HPI-FULL†	29.5	23.9	19.2			35.0
HPI-CAT‡	24.9	23.1	20.9			38.6
TIMI‡	23.2	22.8	20.2			33.0
Eastern Europe:						
HPI-FULL†	23.5	20.5	30.5	14.2		21.5
HPI-CAT‡	18.7	19.9	28.1	12.4		21.7
TIMI‡	16.5	16.1	19.1	9.7		23.3
Russia:						
HPI-FULL†	29.8					
HPI-CAT‡	24.6					
TIMI‡	21.6					
Asia:						
HPI-FULL†	18.8					
HPI-CAT‡	16.3					
TIMI‡	18.2					

*All models calibrated by trial and region.

†Treatment offset incorporated prior to re-calibration.

‡Treatment offset only incorporated for application in GUSTO-I.

In summary the loss of explained variability from using a risk score instead of a model is no worse in any particular region. HPI-CAT generally performed better than TIMI in all regions, except in Asia, however this result is unsubstantiated.

6.3.6.3 Calibration slope

Results for the assessment of calibration slope are shown in Table 6.23. Overall results are not shown as by definition intercept and slope estimates will be 0 and 1 respectively. Also since trial-region level calibration was performed, calibration-in-the-large will be accurate leaving the intercept entirely dependent on the slope. Therefore the intercept is not meaningful and is not presented.

In HERO-2 this analysis is establishing whether a globally based estimate of the relationship between risk scores and mortality applies in individual regions. In the other trials whether a Western based estimate of this relationship holds in non-Western regions is largely being determined.

There were no instances in any region of slope estimates being significantly different from one. Hence trial based estimates of the association between LP or risk index scores with mortality held across regions.

Table 6.23 Calibration slope results for the HERO-2 full model, HPI and TIMI risk scores*

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-3
Western patients:					
HPI-FULL†	1.06 (0.90-1.22)	1.00 (0.94-1.07)	1.01 (0.90-1.12)	1.00 (0.96-1.05)	1.01 (0.95-1.07)
HPI-CAT‡	1.14 (0.96-1.31)	1.00 (0.94-1.06)	0.99 (0.88-1.11)	1.00 (0.96-1.05)	1.00 (0.94-1.06)
TIMI‡	1.07 (0.90-1.24)	1.00 (0.93-1.06)	1.00 (0.89-1.11)	1.00 (0.96-1.05)	0.99 (0.93-1.05)
Latin America:					
HPI-FULL†	1.10 (0.95-1.25)	0.96 (0.74-1.18)	0.88 (0.62-1.15)		1.20 (0.64-1.76)
HPI-CAT‡	1.11 (0.95-1.26)	0.99 (0.77-1.22)	1.01 (0.72-1.30)		1.68 (0.87-2.49)
TIMI‡	1.13 (0.96-1.29)	1.10 (0.85-1.35)	1.02 (0.72-1.32)		1.51 (0.79-2.24)
Eastern Europe:					
HPI-FULL†	0.98 (0.89-1.06)	1.01 (0.65-1.37)	1.38 (0.40-2.36)	0.74 (0.39-1.08)	0.89 (0.69-1.10)
HPI-CAT‡	0.94 (0.86-1.03)	1.03 (0.67-1.39)	1.51 (0.35-2.67)	0.78 (0.40-1.16)	0.97 (0.75-1.19)
TIMI‡	0.93 (0.85-1.02)	0.92 (0.58-1.27)	1.02 (0.20-1.84)	0.74 (0.35-1.12)	1.12 (0.87-1.37)
Russia:					
HPI-FULL†	0.99 (0.92-1.06)				
HPI-CAT‡	0.99 (0.92-1.07)				
TIMI‡	1.00 (0.92-1.08)				
Asia:					
HPI-FULL†	0.91 (0.68-1.13)				
HPI-CAT‡	1.01 (0.76-1.26)				
TIMI‡	1.11 (0.84-1.38)				

*All models calibrated by trial and region.

†Treatment offset incorporated prior to re-calibration.

‡Treatment offset only incorporated for application in GUSTO-I.

In Western countries TIMI performed slightly better than HPI-CAT and as good as HPI-FULL in HERO-2. In the other trials slope estimates were all close to 1 since these patients contributed most of the data.

In other regions there were no consistent patterns to imply an advantage from using a particular risk strategy in a specific region. Sometimes the risk scores performed similarly or results were similar for the HERO-2 strategies. Nevertheless confidence intervals were wide and there is no indication that risk scores, calibrated at the trial-region level, are any less applicable within region subgroups than a full risk model.

6.3.6.4 Calibration plots

The following plots show observed versus predicted mortality according to deciles of predicted risk for the HERO-2 full model and the HPI and TIMI risk scores, all calibrated by trial and region. The solid diagonal line reflects perfect calibration of the model predictions; the dashed lines are fitted smooth nonparametric lowess calibration curves.

The relative frequency histograms demonstrate clearly the reduction in the range of prediction probabilities from using a discrete method of prediction compared to a continuous model. Also note there is more spread between plotted points for the full model compared to the risk scores which means that the full model discriminates better.

In HERO-2, as discussed in Chapter 3, it is evident that HPI performs better than TIMI. There is no visible difference in performance between HPI-CAT and the HERO-2 full model. All approaches over-estimate risk for low-risk patients.

In ASSENT-2 it is clear that the full model does better than the risk scores as its plotted points are much closer to the line of perfect calibration. HPI displayed a similar pattern of miscalibration compared to TIMI but the magnitude of inaccuracy was less.

In ASSENT-3 for patients classified in the lowest risk deciles the risk scores appear to do better than the full model however they are not the same patients and the full model has identified patient groups with lower average mortality compared to the average mortality of patients in the lowest risk score deciles. There is little difference between TIMI and HPI.

In GUSTO-I clearly TIMI performs the worst of all approaches. The full model is very accurate except in the lowest risk patients.

In GUSTO-2 none of the strategies maintained good accuracy. Outside of HERO-2 the performance of the full model deteriorated the most in GUSTO-2.

In GUSTO-3 HPI-FULL performs the best however all approaches do very well with the exception that HPI-CAT over-estimated risk substantially in the second lowest risk decile.

In summary the risk scores do reasonably well compared to the risk model and the relative performance seen in HERO-2 of TIMI and HPI was not noticeably dissimilar in the other trials.

The decrease in accuracy associated with the use of a risk score instead of a model did not appreciably worsen outside of HERO-2.

This analysis is assessing the ability of risk strategies to predict average risk in large groups of homogeneous patients. In smaller groups the HERO-2 full model would display a larger advantage as its predictions are more refined. However based on the *c* statistic and R^2 findings it is unlikely that this advantage would be appreciably larger outside of HERO-2.

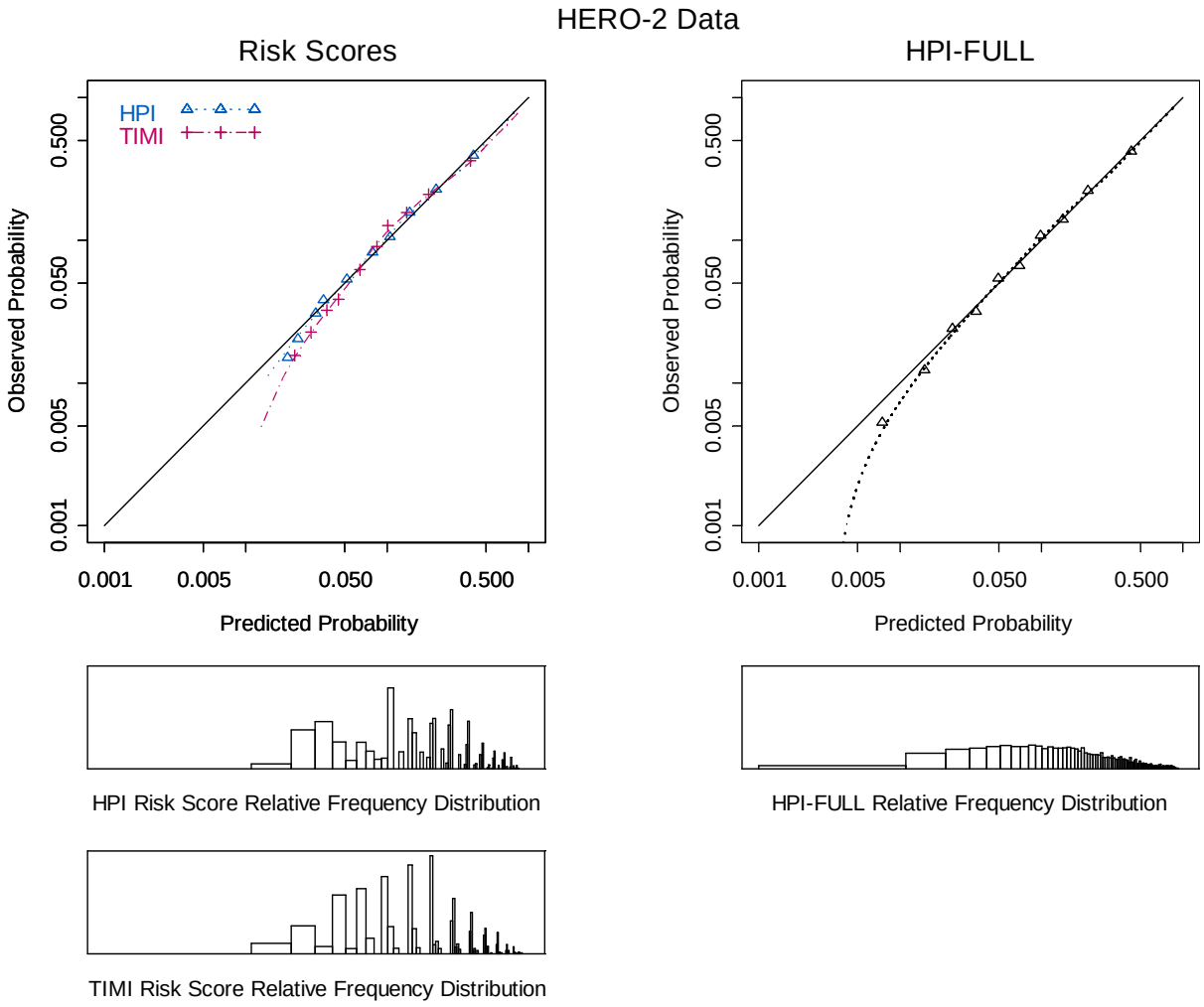

Figure 6.7 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in HERO-2.

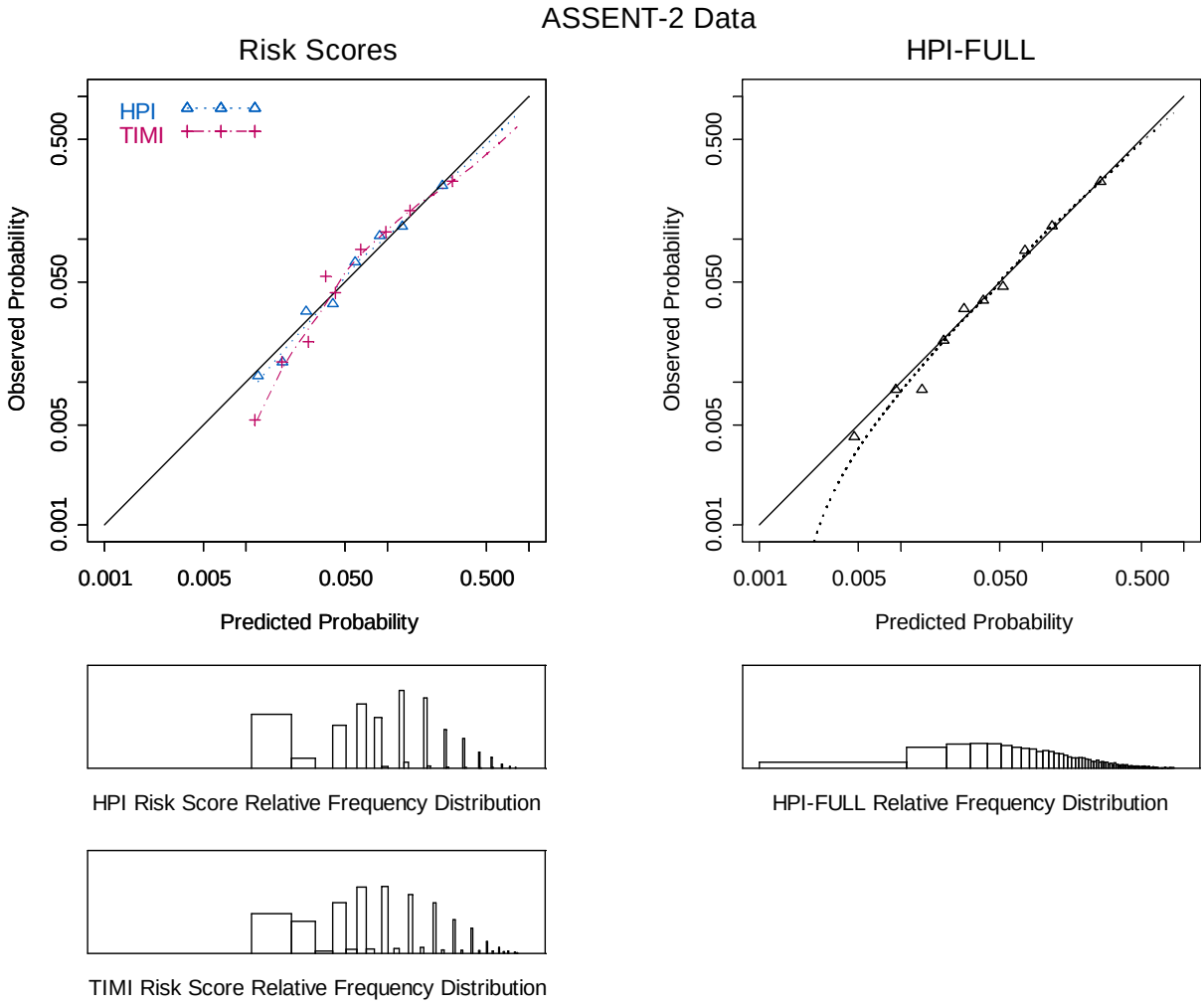

Figure 6.8 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in ASSENT-2.

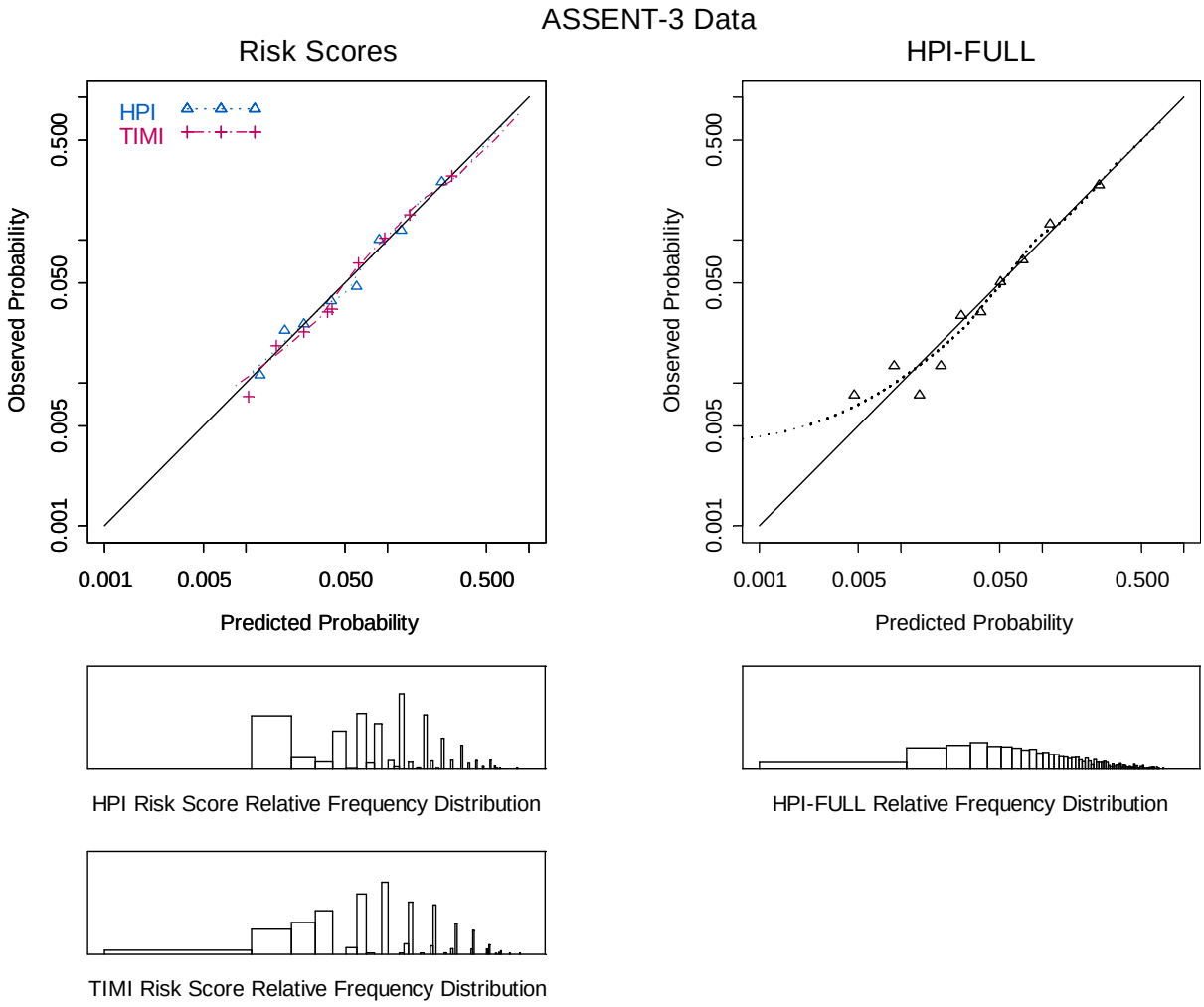

Figure 6.9 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in ASSENT-3.

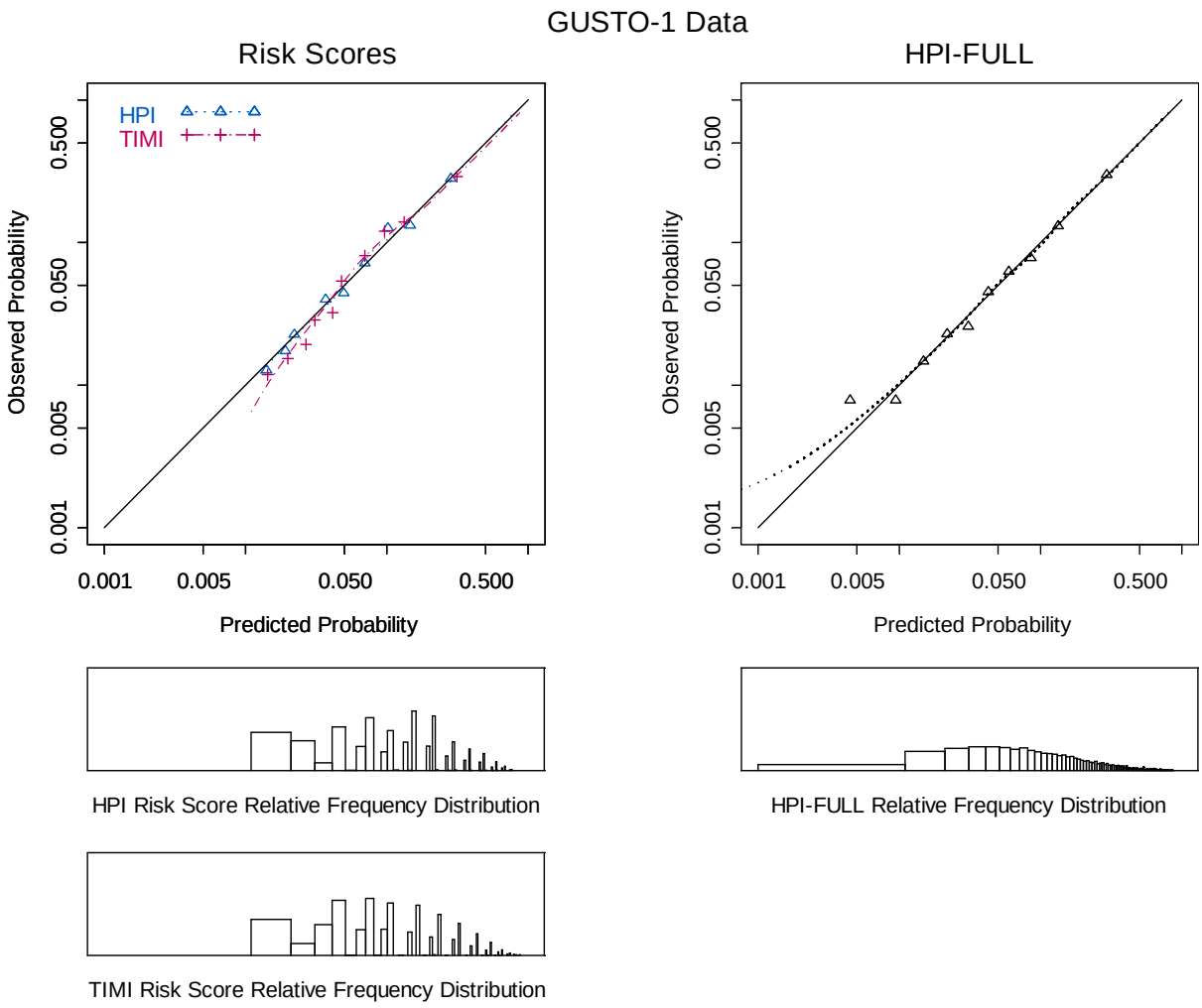

Figure 6.10 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in GUSTO-1.

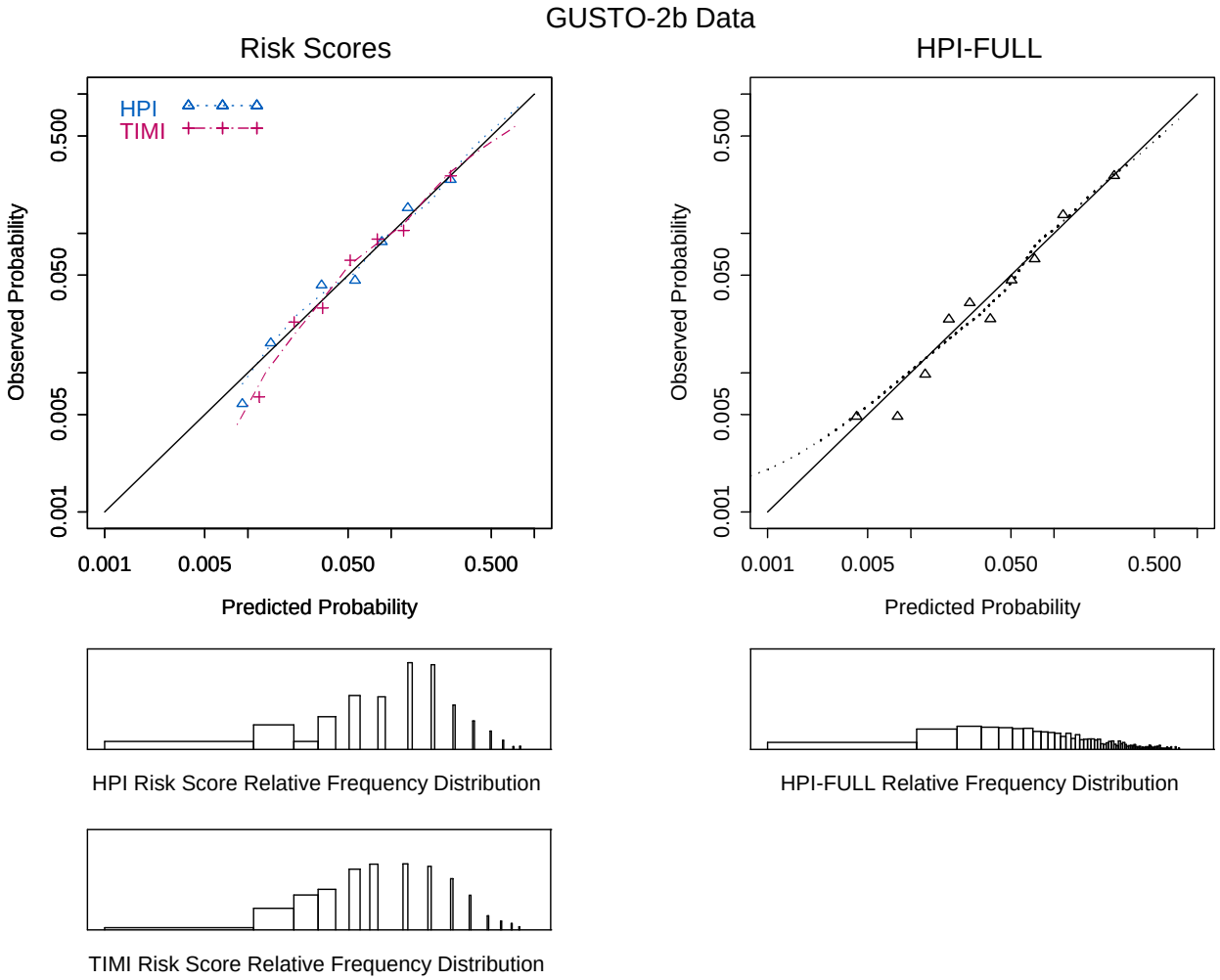

Figure 6.11 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in GUSTO-2b.

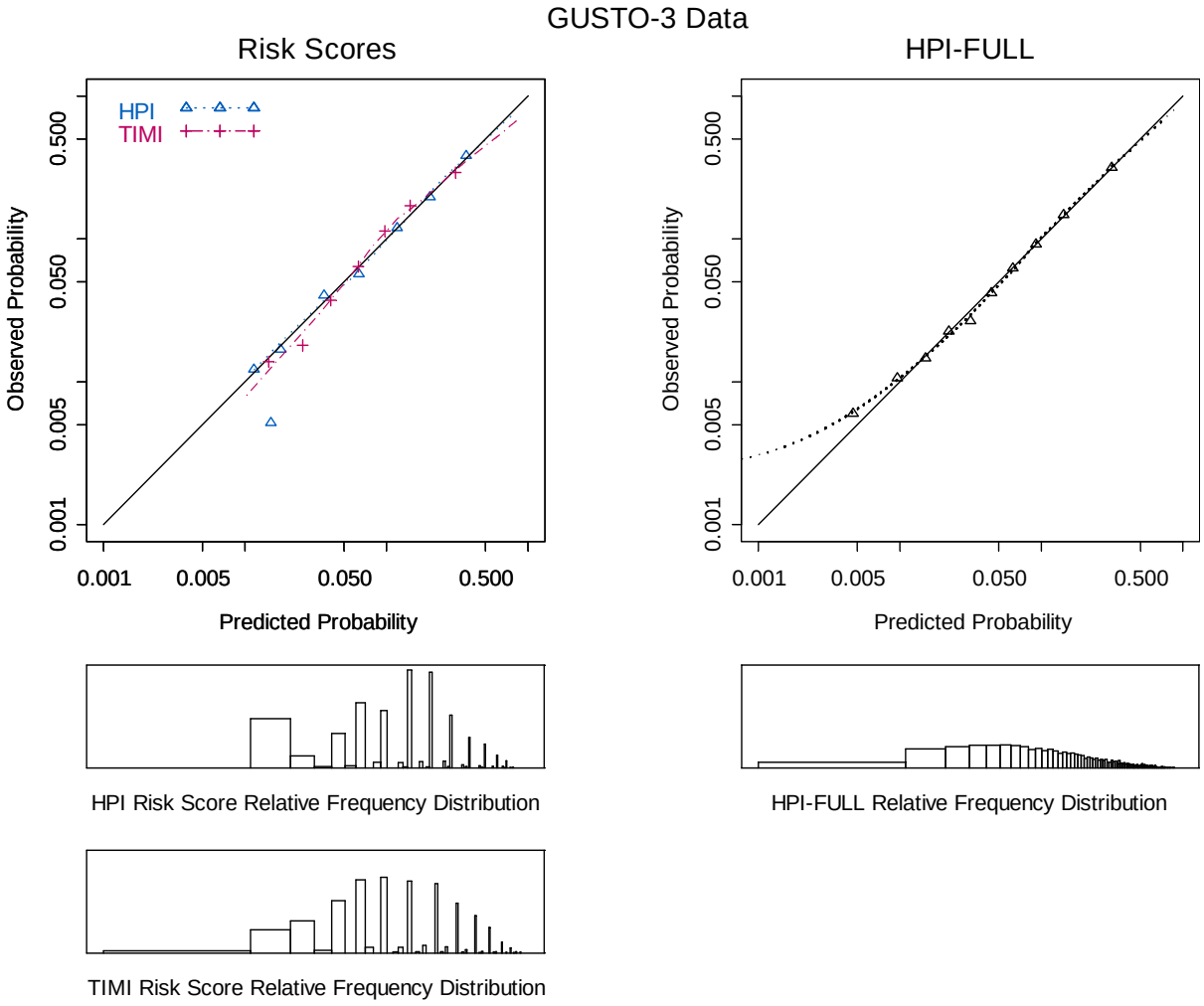

Figure 6.12 HERO-2 full model and the HPI and TIMI risk scores (all calibrated by trial and region) applied in GUSTO-3.

6.4 Conclusion

These results have demonstrated that when applying risk models to a new trial a new model may not need to be developed from scratch and existing models will often suffice. This is with the caveat that simple calibration may be needed to correct for a discrepancy between the average predicted risk and actual risk in the population. Concerning discriminatory ability all of the models are as good as each other: the performance of models applied in another trial was not much worse than the apparent performance of a model developed in that trial; in data independent to all models there was little difference in performance.

Furthermore no model demonstrated an advantage in a particular region. The GUSTO-I model was developed in Western patients but did not perform consistently better in this region. The TIMI risk score, also developed predominantly in Western populations, was not superior to HPI-CAT in this region. The HERO-2 strategies, based largely on Eastern European and Russian patients, did not display a major advantage in these regions. This suggests that new models do not need to be developed for application in other regions, although re-calibration may be needed. All models suffered a decline in performance in Asia, although this is unsubstantiated as only one trial i.e., HERO-2, enrolled patients in this region and it was small.

Previously the HERO-2 strategies had only been validated in Western data using the GUSTO-I database. This VIGOUR collection of trials provided the opportunity to test the HERO-2 models in independent non-Western data. Since these strategies are based on a global sample, performance was, as expected, good.

Many factors may have contributed to the large differences observed in mortality rates across trials. The obvious include differences in patient casemix, patterns of care and available resources. Despite accounting for differences in thrombolytic treatment, re-calibration was still needed. GUSTO-3 Western and Latin American patients had more negative prognostic features on average compared to patients in these regions in other trials. Fittingly the highest mortality for these regions was observed in GUSTO-3 and this trial had the second highest overall mortality. The HERO-2 model captured this increased risk in GUSTO-3 most effectively, with the other models significantly under-estimating the overall rate. Nevertheless there may be additional unmeasured characteristics that contributed to the increased risk.

HERO-2 Russian patients had the highest overall risk profile of all trial-region subgroups as evidenced by the HPI and TIMI risk scores. One reason for this is that a higher proportion of screened patients were randomised in this region, as determined from the HERO-2 screening log substudy (Chapter 4). Consequently this group was more representative of the true patient population in this region.

Even though GUSTO-2b patients were a different group clinically, owing to a less severe eligibility and inclusion criteria, models still performed well after re-calibration.

Initially we applied simple calibration at the trial level to update models. This corrects calibration-in-the-large and adjusts the relationship between risk factors and mortality projected by the model to match the overall relationship in the trial. This was adequate in the regard that no further shrinkage or inflation of regression coefficients was indicated within region subgroups. However calibration-in-the-large within these subgroups was often inaccurate indicating the need for region-specific intercepts, as consistent with earlier chapters, but not region-specific slopes.

The relationship between risk factors and mortality appeared stronger in GUSTO-I, a Western population. This discrepancy is not an artifact of regional differences because there was no consistent evidence that the need for inflation or shrinkage of coefficients is dependent on region (as shown in Table 6.17). Rather updating of coefficients is more a function of trial. These results correspond with the findings of Chapters 3 and 5 where region-specific coefficients were not indicated.

Another important aspect to consider when comparing risk models is the risk factor information that they encapsulate. The GUSTO-I model includes several predictors which were considered for inclusion in other models but were not selected. Despite this the discriminatory ability of these other models was almost as good in GUSTO-I. Comparisons with the ASSENT-3 model showed that supplementing traditional risk factor information with complicated electrocardiographic data provides little improvement. It appears that there is a set of core variables which contain most of the vital information needed to predict survival and that the addition of other variables adds little.

When models were applied to later related trials the size of the advantage displayed over other models in discriminatory ability was small, although not all sites participated in later trials and new sites were activated. We examined the Australian and New Zealand sites that participated in these VIGOUR trials and found overlap in site participation across all six trials further diminishing the value of this comparison. Putting aside these shortcomings, results for the GUSTO-I model relevant to temporal validity were not as hypothesised as shrinkage of model coefficients was indicated in GUSTO-2b; other unrelated models required only intercept adjustment. Conversely the ASSENT-2 model displayed good temporal validity when applied in ASSENT-3 as neither intercept or slope correction was indicated.

Where geographical transportability could be contemplated, good discrimination was displayed by all models. The ASSENT-2 model performed very well in other geographical regions such as Russia and Asia as intercept adjustment was only required. The GUSTO-I model required shrinkage in Eastern Europe and Russia however as stated above there appears to be an underlying trial effect. Hence it is fair to conclude that results relevant to geographical transportability are in line with earlier chapters.

There are many reasons as to why a model may not demonstrate temporal or geographic transportability. In temporal validation, structural differences between datasets might arise from temporal changes; for example revised diagnostic criteria or different referral schemes in the hospital [110], although these factors may be less relevant is this setting. In external or geographical validation even if the same inclusion criteria are employed there are many sources for structural differences

between the two populations including training of the treating physicians, available equipment, cultural and ethnic differences [110]. Geographical regions may deviate with regard to population structure, health-political issues, socio-economic and genetic factors [110]. The external validation considered here additionally included a temporal shift.

As already stated the mis-calibration assessment results suggest the relationship between risk factors and mortality is stronger in GUSTO-I. A major contributor to this would be the stronger age effect in the higher Killip classes in GUSTO-I, i.e., the age by Killip class interaction was weaker. We studied this interaction in Western patients across trials and found the same trend in HERO-2 but not elsewhere. This lacking of consistency is unresolved. Perhaps the availability and utilisation of treatments play a part in how strong the age effect is in the higher Killip classes, i.e., most sick patients. Other differences between models may also relate to patterns of care. Determination of this is beyond the scope of this thesis.

The HERO-2 full model re-estimated excluding angina calibrated better in ASSENT-2 than the original version applied whilst ignoring the angina term. There was no difference in discriminatory ability. After re-calibration the original model was no worse than the re-estimated model in all aspects of performance. This shows that if all of the information required to calculate a model's score is not available, re-calibration can correct any resulting bias.

The ASSENT-3 model demonstrated that race is a predictor of post MI 30-day mortality [9] although region was not considered for inclusion in this model and race is highly correlated with region. ASSENT-3 recruited in Latin America and Eastern Europe in addition to Western countries. In exploratory analysis we found that region is a stronger predictor than race and that adjustment for race adds little information in addition to region. This is consistent with the GUSTO-I model as race was not selected as an independent predictor in Western patients [1].

The risk scores performed well compared to the HERO-2 full model. HPI-FULL has several advantages including an enhanced range of prediction, inclusion of additional prognostic information, and greater refinement as outcome is predicted on a continuous scale (i.e., no rounding is employed). However in some of the smaller subgroups this added precision was detrimental when both ranking patients (*c* statistic) and predicting outcomes (R^2). In GUSTO-3 HPI and TIMI performed better than HPI-FULL in Latin America (n=121) and Eastern Europe (n=886) respectively. The calculation of R^2 involves comparing observed versus expected numbers of events in homogenous risk groups [81]. Numbers in these groups would be smaller for HPI-FULL hence there may be more noise when sample size is small.

The mortality predicted by HPI-FULL and HPI-CAT in their corresponding deciles of lowest risk appeared less accurate for HPI-FULL (i.e., on the log scale) in HERO-2, ASSENT-3 and GUSTO-I. However HPI-FULL is more effective at identifying low risk patients; mortality in the HPI-FULL lowest-risk decile was roughly half the respective rate for HPI-CAT. A cross-tabulation of the HPI-FULL and HPI-CAT deciles of risk groups showed major disagreements in risk classification. If HPI-

FULL is considered the gold standard there is inherent heterogeneity in patients' risk levels within HPI-CAT decile groups.

As stated earlier the HERO-2 mortality rate for Eastern Europe appears to be an outlier as it was much higher than in the other trials for this region. However in pairwise comparisons this increased rate was only significantly higher than ASSENT-2 (P=0.003) and after adjusting for the HERO-2 model predictors the significance was borderline (P=0.052). The use of revascularisation procedures does not further eliminate this difference. Other differences in care or unknown risk factors may have contributed. The huge sample size for Eastern Europe in HERO-2 compared to the other trials should be kept in mind (N=5877).

In Chapter 3 the GUSTO-I model was found to discriminate poorly (c = 0.718) in Asia contradicting findings here. That previous analysis was based only on the HERO-2 heparin treatment arm as the GUSTO-I model precludes treatment with Bivalirudin. Additional analysis showed that all of the models perform this poorly when only applied in the heparin group. In the Bivalirudin group c statistics for all models were around 0.84 and for the two groups combined 0.77. This instability is most likely due to the smaller sample size. The result using both treatment arms should apply.

In conclusion these risk stratification tools are all very consistent in terms of performance across trials and regions. New prediction methods do not need to be developed for application in new trials or other regions. The risk scores are useful and do well compared to the full models, but their limitations should be borne in mind. If precision is integral a full model should be applied. To our knowledge this is the most extensive comparison of risk models ever undertaken based on a database of 90,000 patients. We have compared the performance of four comprehensive multivariable risk models and two risk scores across six large international clinical trials which recruited in diverse geographical areas.

CHAPTER 7: DISCUSSION

7.1 Overall findings

The first major finding is that there was no evidence to suggest the effects of risk factors vary by geographical region, even with substantial information from different regions. This is an important message as despite the disparities in the quality of health care systems around the world and patient differences concerning ethnicity, lifestyle and other factors the effects of traditional risk factors on outcome seem to remain constant. The GUSTO-I, HERO-2 and ASSENT-2 models performed satisfactorily in all regions and no particular model displayed a clear-cut advantage in any particular region. Models developed mainly in Western patients extrapolated to non-Western regions and the HERO-2 model based predominantly on Russia and Eastern Europe held up in other regions. Formal tests of interaction also confirmed no heterogeneity in risk factor effects according to region.

Nevertheless, there is a need to adjust by region. We could not explain regional differences in mortality using patient characteristics, treatments, or national health and economic statistics. Differences in unmeasured genetic factors, ethnic origin or lifestyle, and other non-traditional risk factors may be part of the explanation. It is also possible that our adjustments for patient care were not sufficient to capture real differences in the timing and utilisation of drugs and procedures.

Another major finding is that there was no strong evidence to suggest differential coefficient effects in different trials. This consistency even extends to consistency of interactions between the covariates – such as the interaction between age and Killip class across at least three of the studies. Thus we have a high degree of confidence that the HERO-2 risk strategies and other VIGOUR models can predict in different contexts. There is the caveat that calibration may be required. New models do not need to be developed from scratch. When models were applied to later related trials a slight advantage was displayed suggesting that in clinical trials analyses it may be optimal to choose a model related to the trial at hand for risk stratification, however the effect is slight.

Another important finding of the VIGOUR analysis is that no model developed for other studies was clearly superior. Models performed best in the data set they were developed on but there was not much difference across the trials and in data independent to all models performance was similar. This analysis has provided additional external validation of these risk models.

This work has also demonstrated that inclusion of complicated indices based on ECG data in risk assessment strategies, as in the ASSENT-3 dynamic model, does not provide worthwhile gains in discriminatory performance as demonstrated by the comparison of *c* statistics. Risk assessment based on only traditional risk factor information is adequate.

We have also found that simple risk scores derived by approximating multivariable risk model coefficients do provide good risk stratification; there was some degree of loss of information but performance measures were still very good. For example *c* statistics of approximately 80% were obtained on average for the HERO-2 and TIMI risk scores across the VIGOUR trials.

7.2 Model updating

Calibration is critical when applying risk models. However our analysis in Chapter 5 confirmed Steyerberg's [57] findings that extensive model revision is not worthwhile when applying risk models in other populations which are "plausibily related" and that simple recalibration is adequate. The proviso is that the model was well developed in a large sample and that the validation sample is large enough to judge that the model is useable. Steyerberg recommends that the relevance of the model should be supported by reasonable validity in the sample from the new setting, i.e. some correlation should be present between predictions and outcomes [65]. If this is not the case development of a new model should be considered. In the context of HERO-2, the application of model revision to the global model was not worthwhile in large and small regions; only regional intercept adjustment is needed. This is in harmony with conclusions above. In the context of VIGOUR, simple recalibration of existing models was sufficient to achieve predictive accuracy in new trials.

The degree of benefit from model revision depends on whether there is evidence of interaction in risk factor effects according to population, which in this setting was absent. However, it should be kept in mind that estimates of coefficients in different settings are subject to variability and any differences may just be random variation. One may be estimating coefficients from the same underlying distribution as pointed out by Steyerberg [65]. Obvious reasons as to why the variability in the effects of predictors was small in HERO-2 include: predictors were registered according to uniform definitions, were relatively objective characteristics with limited measurement error, and the quality of the data collection was controlled well in the trial. This pertains also to the VIGOUR investigation (Chapter 6).

Model updating is a formal approach to using prior knowledge. One must be comfortable with assuming that a prior model is valid for a new setting. We preferred not to update or adapt previously developed models and to take full advantage of the high quality HERO-2 data to derive a new AMI risk prediction model. This model is unique as it is based on a geographically diverse sample. We have also avoided the pitfalls of model updating as described by Steyerberg [65]. First, important new predictors may be missed. This was unlikely here as we did not consider potential predictors that have not been tested before. Second, you do not discover new knowledge; only combine it with what is already known. We have used empirical data to learn about patterns in this international population, and to derive a model that can provide predictions for new subjects from this population. Thus the uniqueness of the dataset at hand is another consideration in the decision to derive a new risk model or update a prior strategy.

7.3 Generalisability

Generalisability is a key attribute of a model as it determines how broadly applicable it will be. Generalisability depends on the quality of the prediction model as developed for the development setting (internal validity), and on characteristics of the population where the model is applied (validity

of regression coefficients and distribution of predictor values) [65]. Determinants of external validity related to case-mix include: differences in the distribution of the predictors between the development and validation setting; differences in the distributions of missed predictors, and design differences which affect the outcome and hence may influence case-mix indirectly [65]. The last listed determinant is relevant to GUSTO-2b as the selection criteria for this trial was different, resulting in lower mortality than predicted by risk models. However discriminatory ability was still very good meaning coefficient effects were still applicable. At least for the HERO-2 and ASSENT-2 models, calibration-in-the-large was the only issue as estimates of calibration slope were close to one. The baseline characteristics of patients in GUSTO-2b were similar to the other trials implying there were differences in other characteristics (probably related to the severity of the MI), not measured.

When applying a model developed in a trial population to a broader patient group, extrapolation of model predictions in the validation data (e.g., registry) beyond observed predictor values in the development data should be considered [65]. More heterogeneity in case-mix in the validation data translates into higher discriminatory ability. Conversely a more homogeneous validation sample will result in less discriminatory ability.

Examples from the literature of ACS risk strategies applied in external populations will now be discussed. Recently Rathore et al [112] applied the TIMI risk score to a community-based cohort of almost 50000 elderly patients (aged ≥65 years) admitted with STEMI in US hospitals. It was claimed that the cohort represented the majority of hospitalisations and deaths from STEMI. The performance of TIMI was poor (c = 0.67). There was a near 3-fold difference in 30-day mortality rates between the TIMI derivation cohort (i.e., the study InTIME-II) [3] and the community cohort reflecting the 14-year higher median age and increased comorbidity in the community cohort. The authors concluded that TIMI demonstrated poor spectrum transportability. A more severe case-mix in the validation setting is associated with somewhat less spread in predictions, and hence a lower c statistic as declared above. This would be part of the explanation for the poor discriminatory performance. The study had limitations as patients experienced their MI more than 10 years ago (i.e., 1994-1996) and the treatment received may not have been contemporary; only 37% received reperfusion therapy. The most critical shortcoming was the 65 year age cutoff used which severely reduced the representativeness of the cohort. The proportion of patients suffering MI in general populations aged <65 years is substantial. The average age for a first heart attack is 66 for men and 70 for women in the United States [113]. In the Singh et al study [54] which involved a community registry of patients with MI in Olmsted County, Minnesota, the average age of patients was 67 years.

TIMI's discriminatory ability improved considerably in the Singh study [54], but as mentioned in Chapter 1 was sub-optimal compared to in InTIME-II (c = 0.73 vs. 0.78). In this study patients were older with more risk factors than in InTIME-II and mortality was higher (12 vs. 6.7%). The differences in patient characteristics compared with InTIME-II did not compare with that seen in the Rathore study. In the Lev et al study [23], described in Chapter 1, the performance of TIMI was

similar (c = 0.724). Recall that, the setting in this study was a single-centre registry of patients undergoing PCI. Mortality was lower (3.6%) than in InTIME-II but risk factor rates were not consistently higher or lower. When TIMI was applied to a trial population of patients undergoing PCI its performance deteriorated (c = 0.70) [21]. This is plausible as mortality in this trial was even lower (2.7%). Risk factor rates were lower than in the real-life PCI cohort but generally similar to in InTIME-II. These examples illustrate how differences in patients' risk levels and characteristics, between the development and validation settings, can influence a models performance.

The Cadillac risk score, mentioned in Chapter 1, was derived in a trial population of patients undergoing PCI. This score maintained performance as in its derivation cohort when applied to a real-life cohort (c = 0.83 in derivation vs. 0.82) [21, 23]. The increase in mortality and particularly risk factor rates in going from the Cadillac trial to the real-life cohort was small (30-day mortality: 2.1 vs. 3.6%).

The representativeness of the HERO-2 sample is now considered. Based on a screening log substudy, the level of purported generalisabiltiy varied according to region; patients in Russia and Eastern Europe were more representative as higher proportions of patients entering hospitals were enrolled. Mortality was higher among non-participants than participants, and screened patients were on average 3-4 years older than those randomised. This discrepancy does not rival the 3-fold and 14-yr differences in observed mortality and median age, respectively, between the development (i.e., InTIME-II) and validation settings in the Rathore paper. We anticipate far better translation of the HERO-2 risk strategies to general populations. Further research is required to precisely assess this.

As remarked in Chapter 1 the GRACE risk score performed very poorly when applied to a real-life AMI cohort undergoing PCI [23]. Further elaboration and discussion of this is warranted. GRACE was derived in both STEMI and non-STEMI/UA patients. Lev et al provided a number of possible reasons as to why the score performed so poorly. Firstly, three of its components were irrelevant in their cohort; patients with cardiogenic shock were excluded (cardiac arrest) and all patients had ST-segment deviation and increased cardiac enzymes. Secondly, the score does not include location of infarct, a key risk factor in AMI. The score does encapsulate the key risk factors Killip class, age, heart rate and SYSBP and prior models incorporating just age, heart rate and SYSBP (i.e., Simple Risk Index) have performed well [56]. Perhaps the points scoring scheme is not suitable for AMI. The GRACE developers stated that their risk score was developed to perform equally well in STEMI and NSTEMI and had a high level of generalisability as it was derived using a multinational registry which also allowed variables which are not commonly collected in clinical trials to be identified as important prognosticators [12]. It appears that this score may not be adequate for application in exclusively AMI populations. It would be interesting to know how the full model the risk score was derived from would have performed.

7.4 Simplified risk scores

In this work we have found that whilst there is some loss of information with using a simple risk score in place of a full multivariable model they still perform well. Our findings appear not to accord with some other studies and further examination of this is warranted. Gale et al [56] concluded that simpler risk models performed as well as more complex models when they refitted five ACS models using data from a national registry in England/Wales. However models were not compared across the same patients and outcomes and the simpler models provided predictions on a continuous scale, i.e., they were not discrete scores. It was stated that findings contrasted to those of Yan et al [114] who found that the TIMI risk score for UA/NSTEMI ACS discriminated poorly for in-hospital mortality (c = 0.68) compared to the GRACE (c = 0.81) and PURSUIT scores (c = 0.80) when applied to patients with NSTEMI in a Canadian ACS registry. This TIMI score is very simple containing only 7 dichotomous variables (range 0-7), whereas GRACE (range 1-372) and PURSUIT (range 0-25) use a combination of semi-continuous and categorical variables. The TIMI score was developed to predict a composite endpoint (i.e., death/(re-) infarction or urgent revascularisation through 14 days after randomisation), although it demonstrated better discrimination for death alone in its derivation and validation cohorts (c = 0.65 and 0.61 respectively for composite endpoint vs. c = 0.78 and 0.72 for all-cause mortality) [15]. It appears that too large a tradeoff has been made between simplicity and accuracy. Furthermore the score does not incorporate the hemodynamic variables heart rate and SYSBP.

Another risk score PREDICT (range 0 to 24) [13] was applied in the Singh et al [54] study above and performed well (c = 0.81 for STEMI). The authors concluded that the inclusion of comorbidity in this index was instrumental to its enhanced discrimination compared to the TIMI risk score for STEMI. The scores larger range would also be a factor. PREDICT was developed in a population-based ACS cohort (i.e., UA and AMI) aged 30 to 74 years, assembled in a similar manner to the Olmsted county cohort. Whilst PREDICT is a score, it is rather comprehensive; its components and points scoring scheme fill two A4 pages. Further, it requires more information and is more difficult and timely to compute than the HERO-2 score. The choice of which index to prefer may depend on the desired level of simplicity. The other benefits of HPI are that it was based on a purely AMI population with no age constraint.

7.5 Other issues

The feasibility and usefulness of risk assessment tools in AMI for clinical practice will now be addressed. Regression formulas can be programmed into mobile devices, electronic patient records etc, however entering the required input data takes time and consequently the feasibility of using risk assessment strategies in the acute phase of MI has been questioned. The clinical usefulness of indexes with regard to aiding in treatment decisions has also been questioned. In an editorial in the American Heart Journal [115] Veronique Roger stated "*the treatment of STEMI is largely standardised and based on evidence from large-scale reperfusion clinical trials, such that little change in management*

can be expected to arise from such scoring systems". Despite this concern nevertheless there has been considerable development in the literature. Kent et al [11] very recently developed a mathematical model that predicts mortality in patients with STEMI if treated with PCI and if treated with thrombolytic therapy. The model, labeled PCI-TPI, was adapted for incorporation into a conventional computerised ECG, with the purpose of supporting physicians considering the risks and benefits of PCI with some delay compared with more immediate thrombolytic therapy. The highest risk tertile of patients accounted for virtually all the mortality benefit from PCI, consistent with earlier work [116]. The authors considered that in clinical practice, if installed in ECGs, the PCI-TPI may be useful in targeting PCI to patients who might benefit most, especially in circumstances where there were substantial barriers in implementing a strategy of PCI for all. Despite the growing evidence that angioplasty yields better outcomes, thrombolytic therapy remains the most common form of reperfusion therapy in AMI because of limited capacity for PCI at most hospitals as asserted by Kent et al [116].

The recent clinical trial CARESS-in-AMI [117] randomised patients with STEMI treated by thrombolytics and abciximab at a non-interventional hospital to immediate transfer for PCI, or to standard medical therapy with transfer for rescue angioplasty. This illustrates that the use of PCI is a current treatment dilemma in STEMI and provides a situation in which risk assessment tools would be clinically useful. The trial showed that immediate transfer for PCI improves outcomes in high-risk patients.

In a study by Yan et al [114] in which risk scores (i.e., TIMI, GRACE and PURSUIT) were applied to patients with NSTE ACS, study physicians were asked to rate patients risk as low, moderate or high on the case report form. The risk scores conferred independent and greater prognostic information compared with physicians' risk assessment for prediction of 1-year mortality, even after adjusting for treatment differences.

Given the difficulties noted above risk assessment strategies may find appropriate use more frequently during the later period of the hospitalisation or for prediction of longer-term outcomes. Other important uses include case-mix adjustment, and risk stratification in clinical trials analyses and inter-hospital comparisons. In provider profiling analysis hospitals are ranked according to performance. This analysis plays a role in quality improvement; hospitals with a poor ranking are given an incentive to review their processes of care delivery and improve. Case-mix adjustment is very important, as some hospitals may treat more severe patients; this confounding needs adjustment [65].

We have focused our evaluation on the endpoint of 30-day mortality. In the editorial mentioned above Roger [115] also asserted that more focus should be placed on longer-term outcomes following AMI, such as HF and recurrent ischemic events, "as these outcomes have not declined over time in ways commensurate to the large declines in short-term case fatality rates observed in clinical trials". We did not have long-term data available. Whether our findings extrapolate to long-term outcomes including survival is another line of research.

Biomarkers have many uses. They can be used to categorise disease severity, to guide the use of targeted therapies, to monitor the efficacy of therapy and to predict disease course, including recurrence and response to therapy [118]. However the use of many biomarkers for rapid risk assessment in the emergency department or at the bedside is infeasible as this information is often not immediately available [118]. Perhaps biomarkers should be concentrated more towards the prediction of long-term outcomes. For example a tool has been developed called the TIMI HF risk score which assesses the risk of developing post-ACS HF leading to hospitalisation [119]. The score, developed in UA or NSTEMI, can accurately predict patients at highest risk for the development of HF up to 10 months following the initial event. The score incorporates basic clinical findings as well as a biomarker BNP which enhanced its discriminatory capacity. However, the inclusion of biomarkers as prognosticators in risk models will preclude direct application when this data is not available. If collection of such data were to become standard practice, the applicability of the findings here to biomarkers would need to be examined.

The variability in external validation performance measures depends on the size of the development and validation samples. The large trial datasets in this study have afforded large power to detect invalidity in prediction models and provided reliable estimates of performance measures. This is a major strength of this research.

7.6 Summary

We have developed a multivariable risk model based on a geographically diverse trial population and simplified this to form a reduced model and a simple risk score. Our risk strategies may be preferred for use in geographically diverse samples. The HERO-2 reduced model miscalibrated in GUSTO-I due to omitted factors which were underrepresented in HERO-2. Simple recalibration addressed this and is recommended even within the same geographical region. The HPI risk score performed a little better than TIMI most of the time. There is some loss of information associated with using discrete risk scores, but they still provided very good risk assessment in this study. In application, it is important to contemplate the suitability of existing risk scores for the patient population in the new setting.

REFERENCES

1. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-I Investigators. Circulation 1995; 91: 1659-68.
2. Califf RM, Woodlief LH, Harrell FE, Lee KL, White HD, Guerci A, et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 1997; 133: 630-9.
3. Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, et al. TIMI risk score for ST-elevation myocardial infarction: A convenient, bedside, clinical score for risk assessment at presentation: An intravenous nPA for treatment of infarcting myocardium early II trial substudy. Circulation 2000; 102: 2031-7.
4. Maggioni AP, Maseri A, Fresco C, Franzosi MG, Mauri F, Santoro E, et al. Age-related increase in mortality among patients with first myocardial infarctions treated with thrombolysis. The Investigators of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2). N Engl J Med 1993; 329: 1442-8.
5. Morrow DA, Antman EM, Giugliano RP, Cairns R, Charlesworth A, Murphy SA, et al. A simple risk index for rapid initial triage of patients with ST-elevation myocardial infarction: an InTIME II substudy. Lancet 2001; 358: 1571-5.
6. Califf RM, Pieper KS, Lee KL, Van De Werf F, Simes RJ, Armstrong PW, et al. Prediction of 1-year survival after thrombolysis for acute myocardial infarction in the global utilization of streptokinase and TPA for occluded coronary arteries trial. Circulation 2000; 101: 2231-8.
7. Krumholz HM, Chen J, Chen YT, Wang Y, Radford MJ. Predicting one-year mortality among elderly survivors of hospitalization for an acute myocardial infarction: results from the Cooperative Cardiovascular Project. J Am Coll Cardiol 2001; 38: 453-9.
8. Mueller HS, Cohen LS, Braunwald E, Forman S, Feit F, Ross A, et al. Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. Analyses of patient subgroups in the Thrombolysis in Myocardial Infarction (TIMI) trial, phase II. Circulation 1992; 85: 1254-64.
9. Chang W-C, Kaul P, Fu Y, Westerhout CM, Granger CB, Mahaffey KW, et al. Forecasting mortality: dynamic assessment of risk in ST-segment elevation acute myocardial infarction. Eur Heart J 2006; 27: 419-26.
10. Selker HP, Griffith JL, Beshansky JR, Schmid CH, Califf RM, D'Agostino RB, et al. Patient-specific predictions of outcomes in myocardial infarction for real-time emergency use: a thrombolytic predictive instrument. Ann Intern Med 1997; 127: 538-56.
11. Kent DM, Ruthazer R, Griffith JL, Beshansky JR, Concannon TW, Aversano T, et al. A percutaneous coronary intervention-thrombolytic predictive instrument to assist choosing between immediate thrombolytic therapy versus delayed primary percutaneous coronary intervention for acute myocardial infarction. Am J Cardiol 2008; 101: 790-5.
12. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, et al. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med 2003; 163: 2345-53.
13. Jacobs DR, Kroenke C, Crow R, Deshpande M, Gu DF, Gatewood L, et al. PREDICT: A simple risk score for clinical severity and long-term prognosis after hospitalization for acute myocardial infarction or unstable angina: the Minnesota heart survey. Circulation 1999; 100: 599-607.
14. Boersma E, Pieper KS, Steyerberg EW, Wilcox RG, Chang WC, Lee KL, et al. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. The PURSUIT Investigators. Circulation 2000; 101: 2557-67.
15. Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, et al. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. Jama 2000; 284: 835-42.
16. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, et al. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. Jama 2004; 291: 2727-33.
17. Giraldez RR, Sabatine MS, Morrow DA, Mohanavelu S, McCabe CH, Antman EM, et al. Baseline hemoglobin concentration and creatinine clearance composite laboratory index improves risk stratification in ST-elevation myocardial infarction. Am Heart J 2009; 157: 517-24.
18. Woo KS. Perusal of risk stratification of acute myocardial infarction for half a century. Eur Heart J 2009; 30: 1030-2.
19. Kempf T, Bjorklund E, Olofsson S, Lindahl B, Allhoff T, Peter T, et al. Growth-differentiation factor-15 improves risk stratification in ST-segment elevation myocardial infarction. Eur Heart J 2007; 28: 2858-65.
20. Keeley EC, Boura JA, Grines CL. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. Lancet 2003; 361: 13-20.

21. Halkin A, Singh M, Nikolsky E, Grines CL, Tcheng JE, Garcia E, et al. Prediction of mortality after primary percutaneous coronary intervention for acute myocardial infarction: the CADILLAC risk score. J Am Coll Cardiol 2005; 45: 1397-405.
22. Addala S, Grines CL, Dixon SR, Stone GW, Boura JA, Ochoa AB, et al. Predicting mortality in patients with ST-elevation myocardial infarction treated with primary percutaneous coronary intervention (PAMI risk score). Am J Cardiol 2004; 93: 629-32.
23. Lev EI, Kornowski R, Vaknin-Assa H, Porter A, Teplitsky I, Ben-Dor I, et al. Comparison of the predictive value of four different risk scores for outcomes of patients with ST-elevation acute myocardial infarction undergoing primary percutaneous coronary intervention. Am J Cardiol 2008; 102: 6-11.
24. Brener SJ, Westerhout CM, Fu Y, Todaro TG, Moliterno DJ, Wagner GS, et al. Contribution of angiographic and electrocardiographic parameters of reperfusion to prediction of mortality and morbidity after acute ST-elevation myocardial infarction: Insights from the Assessment of Pexelizumab in Acute Myocardial Infarction trial. Am Heart J 2009; 158: 755-60.
25. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
26. Augustin N, Sauerbrei W, Schumacher M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. Statistical Modelling 2005; 5: 95-118.
27. Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and regression trees. Belmont, Calif.: Wadsworth; 1984.
28. Guerriere MR, Detsky AS. Neural networks: what are they? Ann Intern Med 1991; 115: 906-7.
29. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer 2001; 91: 1636-42.
30. Negassa A, Monrad ES, Bang JY, Srinivas VS. Tree-structured risk stratification of in-hospital mortality after percutaneous coronary intervention for acute myocardial infarction: a report from the New York State percutaneous coronary intervention database. Am Heart J 2007; 154: 322-9.
31. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998; 17: 2501-8.
32. National Heart Foundation of Australia. Heart Information. 2010 [updated October 2009; cited May 2010]; Available from: http://www.heartfoundation.org.au.
33. Mackay J, Mensah G, (eds). The atlas of heart disease and stroke. Geneva: World Health Organization; 2004 [cited June 2008]; Available from: http://www.who.int/cardiovascular_diseases/resources/atlas/en/index.html.
34. Leeder S, Raymond S, Greenberg H, Hui L, Esson K. A race against time. The challenge of cardiovascular disease in developing economies. New York: Columbia University 2004.
35. Wilhelmsen L, Welin L, Svardsudd K, Wedel H, Eriksson H, Hansson PO, et al. Secular changes in cardiovascular risk factors and attack rate of myocardial infarction among men aged 50 in Gothenburg, Sweden. Accurate prediction using risk models. J Intern Med 2008; 263: 636-43.
36. InTIME-II Investigators. Intravenous NPA for the treatment of infarcting myocardium early; InTIME-II, a double-blind comparison of single-bolus lanoteplase vs accelerated alteplase for the treatment of patients with acute myocardial infarction. Eur Heart J 2000; 21: 2005-13.
37. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993; 329: 673-82.
38. The Assessment of the Safety and Efficacy of a New Thrombolytic Regimen (ASSENT)-3 Investigators. Efficacy and safety of tenecteplase in combination with enoxaparin, abciximab, or unfractionated heparin: the ASSENT-3 randomised trial in acute myocardial infarction. Lancet 2001; 358: 605-13.
39. The Hirulog and Early Reperfusion or Occlusion (HERO)-2 Trial Investigators. Thrombin-specific anticoagulation with bivalirudin versus heparin in patients receiving fibrinolytic therapy for acute myocardial infarction: the HERO-2 randomised trial. Lancet 2001; 358: 1855-63.
40. Shen JJ, Wan TT, Perlin JB. An exploration of the complex relationship of socioecologic factors in the treatment and outcomes of acute myocardial infarction in disadvantaged populations. Health Serv Res 2001; 36: 711-32.
41. Alter DA, Naylor CD, Austin P, Tu JV. Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. N Engl J Med 1999; 341: 1359-67.
42. Kaplan GA, Keil JE. Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation 1993; 88: 1973-98.
43. Blumenthal JA, Babyak MA, Carney RM, Huber M, Saab PG, Burg MM, et al. Exercise, depression, and mortality after myocardial infarction in the ENRICHD trial. Med Sci Sports Exerc 2004; 36: 746-55.
44. Abete P, Ferrara N, Cacciatore F, Sagnelli E, Manzi M, Carnovale V, et al. High level of physical activity preserves the cardioprotective effect of preinfarction angina in elderly patients. J Am Coll Cardiol 2001; 38: 1357-65.

45. Hedback B, Perk J, Wodlin P. Long-term reduction of cardiac mortality after myocardial infarction: 10-year results of a comprehensive rehabilitation programme. Eur Heart J 1993; 14: 831-5.
46. Janszky I, Ljung R, Ahnve S, Hallqvist J, Bennet AM, Mukamal KJ. Alcohol and long-term prognosis after a first acute myocardial infarction: the SHEEP study. Eur Heart J 2008; 29: 45-53.
47. Mukamal KJ, Maclure M, Muller JE, Mittleman MA. Binge drinking and mortality after acute myocardial infarction. Circulation 2005; 112: 3839-45.
48. Kromhout D, Menotti A, Kesteloot H, Sans S. Prevention of coronary heart disease by diet and lifestyle: evidence from prospective cross-cultural, cohort, and intervention studies. Circulation 2002; 105: 893-8.
49. Gupta M, Chang W-C, Van de Werf F, Granger CB, Midodzi W, Barbash G, et al. International differences in in-hospital revascularization and outcomes following acute myocardial infarction: a multilevel analysis of patients in ASSENT-2. Eur Heart J 2003; 24: 1640-50.
50. Austin PC, Tu JV, Alter DA. Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? Am Heart J 2003; 145: 27-35.
51. Chang W-C, Midodzi WK, Westerhout CM, Boersma E, Cooper J, Barnathan ES, et al. Are international differences in the outcomes of acute coronary syndromes apparent or real? A multilevel analysis. J Epidemiol Community Health 2005; 59: 427-33.
52. Giugliano RP, Llevadot J, Wilcox RG, Gurfinkel EP, McCabe CH, Charlesworth A, et al. Geographic variation in patient and hospital characteristics, management, and clinical outcomes in ST-elevation myocardial infarction treated with fibrinolysis. Results from InTIME-II. Eur Heart J 2001; 22: 1702-15.
53. Earl-Slater A. Critical appraisal of clinical trials. Advantages and disadvantages of evidence from clinical trials. British Journal of Clinical Governance 2001; 6: 136-9.
54. Singh M, Reeder GS, Jacobsen SJ, Weston S, Killian J, Roger VL. Scores for post-myocardial infarction risk stratification in the community. Circulation 2002; 106: 2309-14.
55. Rathore SS, Weinfurt KP, Gross CP, Krumholz HM. Validity of a simple ST-elevation acute myocardial infarction risk index: are randomized trial prognostic estimates generalizable to elderly patients? Circulation 2003; 107: 811-6.
56. Gale CP, Manda SOM, Weston CF, Birkhead JS, Batin PD, Hall AS. Evaluation of risk scores for risk stratification of acute coronary syndromes in the Myocardial Infarction National Audit Project (MINAP) database. Heart 2009; 95: 221-7.
57. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23: 2567-86.
58. Yan AT, Jong P, Yan RT, Tan M, Fitchett D, Chow C-M, et al. Clinical trial-derived risk model may not generalize to real-world patients with acute coronary syndrome. Am Heart J 2004; 148: 1020-7.
59. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 1988; 7: 257-74.
60. Van De Werf F, Adgey J, Ardissino D, Armstrong PW, Aylward P, Barbash G, et al. Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. Lancet 1999; 354: 716-22.
61. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO) IIb investigators. A comparison of recombinant hirudin with heparin for the treatment of acute coronary syndromes. N Engl J Med 1996; 335: 775-82.
62. The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO III) Investigators. A comparison of reteplase with alteplase for acute myocardial infarction. N Engl J Med 1997; 337: 1118-23.
63. Chambers JM, Hastie T, eds. Statistical models in S. Pacific Grove, CA: Wadsworth & Brooks/Cole; 1992.
64. Hastie T, Tibshirani R. Generalized additive models. Monographs on statistics and applied probability. London; New York: Chapman and Hall; 1990.
65. Steyerberg EW. Clinical prediction models: a practical approach to model development, validation, and updating. New York; London: Springer; 2009.
66. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54: 774-81.
67. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.
68. Alzola C, Harrell F. An introduction to S and the Hmisc and Design libraries 2006. Available from: http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RS.
69. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Medical decision making 1993; 13: 49-58.
70. Cox DR. Two further applications of a model for binary regression. Biometrika 1958; 45: 562–5.

71. Nagelkerke NJD. A note on the general definition of the coefficient of determination. Biometrika 1991; 78: 691-2.
72. Arkes HR, Dawson NV, Speroff T, Harrell FE, Alzola C, Phillips R, et al. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. SUPPORT Investigators. Medical decision making 1995; 15: 120-31.
73. Efron B, Tibshirani R. An introduction to the bootstrap. Monographs on statistics and applied probability. New York: Chapman & Hall; 1993.
74. Miller AJ. Subset selection in regression. Monographs on Statistics and Applied Probability, Vol. 40. London: Chapman & Hall; 1990.
75. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006; 25: 127-41.
76. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994; 86: 829-35.
77. Schumacher M, Hollander N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? Stat Med 1997; 16: 2813-27.
78. Hollander N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. Stat Med 2004; 23: 1701-13.
79. Akkerhuis KM, Deckers JW, Boersma E, Harrington RA, Stepinska J, Mahaffey KW, et al. Geographic variability in outcomes within an international trial of glycoprotein IIb/IIIa inhibition in patients with acute coronary syndromes. Results from PURSUIT. Eur Heart J 2000; 21: 371-81.
80. Domanski M, Antman EM, McKinlay S, Varshavsky S, Platonov P, Assmann SF, et al. Geographic variability in patient characteristics, treatment and outcome in an International Trial of Magnesium in acute myocardial infarction. Control Clin Trials 2004; 25: 553-62.
81. Collett D. Modelling binary data. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2003.
82. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837-45.
83. Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157-72; discussion 207-12.
84. Pregibon D. Logistic Regression Diagnostics. The Annals of Statistics 1981; 9: 705-24.
85. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
86. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). Jama 2001; 285: 2486-97.
87. Fox KAA, Goodman SG, Klein W, Brieger D, Steg PG, Dabbous O, et al. Management of acute coronary syndromes. Variations in practice and outcome; findings from the Global Registry of Acute Coronary Events (GRACE). Eur Heart J 2002; 23: 1177-89.
88. Wong CK, White HD. Has the mortality rate from acute myocardial infarction fallen substantially in recent years? Eur Heart J 2002; 23: 689-92.
89. Kuch B, Bolte HD, Hoermann A, Meisinger C, Loewel H. What is the real hospital mortality from acute myocardial infarction? Epidemiological vs clinical view. Eur Heart J 2002; 23: 714-20.
90. Gaziano TA. Cardiovascular disease in the developing world and its cost-effective management. Circulation 2005; 112: 3547-53.
91. Yusuf S, Mehta SR, Xie C, Ahmed RJ, Xavier D, Pais P, et al. Effects of reviparin, a low-molecular-weight heparin, on mortality, reinfarction, and strokes in patients with acute myocardial infarction presenting with ST-segment elevation. Jama 2005; 293: 427-35.
92. Antman EM, Morrow DA, McCabe CH, Murphy SA, Ruda M, Sadowski Z, et al. Enoxaparin versus unfractionated heparin with fibrinolysis for ST-elevation myocardial infarction. N Engl J Med 2006; 354: 1477-88.
93. Yusuf S, Mehta SR, Chrolavicius S, Afzal R, Pogue J, Granger CB, et al. Effects of fondaparinux on mortality and reinfarction in patients with acute ST-segment elevation myocardial infarction: the OASIS-6 randomized trial. Jama 2006; 295: 1519-30.
94. Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: Part II: variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. Circulation 2001; 104: 2855-64.
95. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006; 3: e442.
96. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. Curr Probl Cardiol 2010; 35: 72-115.
97. Boersma E, Mercado N, Poldermans D, Gardien M, Vos J, Simoons ML. Acute myocardial infarction. Lancet 2003; 361: 847-58.
98. McDermott MM. The international pandemic of chronic cardiovascular disease. Jama 2007; 297: 1197-206.

99. Orlandini A, Diaz R, Wojdyla D, Pieper K, Van de Werf F, Granger CB, et al. Outcomes of patients in clinical trials with ST-segment elevation myocardial infarction among countries with different gross national incomes. Eur Heart J 2006; 27: 527-33.
100. The World Bank Group. The World Bank. ; 2002 [cited June 20 2002]; Available from: www.worldbank.org.
101. World Health Report 2000. Statistical annex. Geneva: World Health Organization 2001.
102. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. Canadian journal of public health 2001; 92: 150-4.
103. Llevadot J, Giugliano RP, Antman EM, Wilcox RG, Gurfinkel EP, Henry T, et al. Availability of on-site catheterization and clinical outcomes in patients receiving fibrinolysis for ST-elevation myocardial infarction. Eur Heart J 2001; 22: 2104-15.
104. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven; 1998.
105. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130: 515-24.
106. Moons KGM, Donders ART, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 2004; 57: 1262-70.
107. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990 Nov; 9: 1303-25.
108. Duke Clinical Research Institute. Virtual Coordinating Center for Global Collaborative Cardiovascular Research. 1997-1998 [cited May 2010]; Available from: http://vigour.dcri.duke.edu
109. Hindman NB, Schocken DD, Widmann M, Anderson WD, White RD, Leggett S, et al. Evaluation of a QRS scoring system for estimating myocardial infarct size. V. Specificity and method of application of the complete system. Am J Cardiol 1985; 55: 1485-90.
110. Konig IR, Malley JD, Weimar C, Diener HC, Ziegler A. Practical experiences on the necessity of external validation. Stat Med 2007; 26: 5499-511.
111. Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet 1986; 1: 397-402.
112. Rathore SS, Weinfurt KP, Foody JM, Krumholz HM. Performance of the Thrombolysis in Myocardial Infarction (TIMI) ST-elevation myocardial infarction risk score in a national cohort of elderly patients. Am Heart J 2005; 150: 402-10.
113. Woolston C. What are your odds of a heart attack? : Cable News Network; 2008 [cited May 2010]; Available from: http://edition.cnn.com/2008/HEALTH/conditions/07/14/healthmag.heartattack.odds/index.html.
114. Yan AT, Yan RT, Tan M, Casanova A, Labinaz M, Sridhar K, et al. Risk scores for risk stratification in acute coronary syndromes: useful but simpler is not necessarily better. Eur Heart J 2007; 28: 1072-8.
115. Roger VL. To score or not to score? Am Heart J 2005; 150: 371-2.
116. Kent DM, Schmid CH, Lau J, Selker HP. Is primary angioplasty for some as good as primary angioplasty for all? J Gen Intern Med 2002; 17: 887-94.
117. Di Mario C, Dudek D, Piscione F, Mielecki W, Savonitto S, Murena E, et al. Immediate angioplasty versus standard therapy with rescue angioplasty after thrombolysis in the Combined Abciximab REteplase Stent Study in Acute Myocardial Infarction (CARESS-in-AMI): an open, prospective, randomised, multicentre trial. Lancet 2008; 371: 559-68.
118. Cannon CP, Greenberg BH. Risk stratification and prognostic factors in the post-myocardial infarction patient. Am J Cardiol 2008; 102: 13G-20G.
119. Wylie JV, Murphy SA, Morrow DA, de Lemos JA, Antman EM, Cannon CP. Validated risk score predicts the development of congestive heart failure after presentation with unstable angina or non-ST-elevation myocardial infarction: results from OPUS-TIMI 16 and TACTICS-TIMI 18. Am Heart J 2004; 148: 173-80.

APPENDICES

APPENDIX A

Example Splus code for shrinkage of region-specific regression coefficients in Asia

##Fit Global model†

```
HGR<-lrm(DEATH~REGION+AGESTD*KILGRP+BPSYS+PULT+SEX+ANT,maxit = 25,data=all, x=T,
y=T, linear.predictors=T)
```

##Copy linear predictor to dataset

```
all$lpg<-HGR$linear.predictors
```

##Estimate adjustments to global regression coefficients to derive revised coefficients for application in Asia

```
AsiaLP<-lrm(DEATH~AGESTD*KILGRP+BPSYS+PULT+SEX+ANT+offset(lpg),
subset=c(REGION==5), data=all, maxit=25, x=T,y=T)
```

##Apply shrinkage – adjustments are shrunk towards zero

```
Asiapen<-pentrace(AsiaLP, 1, method='optimize', maxit=25)

Asiashrunk <- update(AsiaLP, penalty=Asiapen$penalty, maxit=25)
```

##Create matrix to store updated coefficients for Asia

```
Asiafinal<-
matrix(nrow=1,ncol=10,dimnames=list(NULL,c("Intercept","AGESTD","KILLIP2","KILLIP34
","SYSTOLIC","PULT","SEX","ANT","AGESTD.KILLIP2","AGESTD.KILLIP34")))
```

##Calculate region-specific shrunk coefficients for Asia

```
Asiafinal["Intercept"] <-
HGR$coefficients["Intercept"]+HGR$coefficients["REGION=5"]+Asiashrunk$coefficients[
"Intercept"]

Asiafinal[2:10]<- HGR$coefficients[6:14] + Asiashrunk$coefficients[2:10]
```

†KILGRP=3 includes Killip classes 3 and 4.

APPENDIX B

TABLE A1: Trial characteristics

Trial	Sample size	No. hospitals	No. countries	Trialing	Randomised treatments	Additional therapy
GUSTO-1	41021	1081	15	Fibrinolytic/ Antithrombotic therapies	SK + SC heparin **vs** SK + IV heparin **vs** accelerated t-PA + IV heparin **vs** t-PA + SK + IV heparin	aspirin
GUSTO-2b	12142	373	13	Antithrombotic therapies	hirudin **vs** heparin	aspirin, use of SK/accelerated t-PA in STE decision of treating physician
GUSTO-3	15059	807	20	Fibrinolytic therapies	reteplase **vs** accelerated alteplase	aspirin, heparin
HERO-2	17073	539	46	Antithrombotic therapies	bivalirudin **vs** heparin	streptokinase, aspirin
ASSENT-2	16949	1021	29	Fibrinolytic therapies	tenecteplase **vs** alteplase	aspirin, heparin
ASSENT-3	6095	575	26	Antithrombotic therapies	Enoxaparin (max 7 days) **vs** low dose unfractionated heparin ± abciximab	tenecteplase (half or full dose), aspirin

	Recruitment	Blinding	Patients	Trial design	Primary Outcome
GUSTO-1	DEC90-FEB93	Open label	AMI	Efficacy	Death 30 d
GUSTO-2b	MAY94-OCT95	Double blind	Unstable angina/non Q-wave AMI	Efficacy	Death/ReMI/MI 30 d
GUSTO-3	OCT95-JAN97	Open label	AMI	Efficacy	Death 30 d
HERO-2	NOV98 – MAY01	Open label	AMI	Efficacy	Death 30 d
ASSENT-2	OCT97-NOV98	Double blind	AMI	Equivalence	Death 30 d
ASSENT-3	MAY2000-APR2001	Open label	AMI	Non-inferiority	Death 30 d/in-hospital ReMI or refractory ischaemia

	Other Outcomes	Eligibility, inclusion criteria.	Primary analysis	Adjudicated endpoints	Results
GUSTO-1	Death/stroke, death/ICH, death/disabling stroke, death 24 h, stroke and bleeding.	Onset symptoms within 6 h, chest pain lasting ≥ 20 m and STE of ≥ 0.1 mV in ≥ 2 limb leads **OR** ≥ 0.2 mV in ≥ 2 contiguous precordial leads.	ITT	Stroke	Accelerated t-PA significantly better survival + NCB.
GUSTO-2b	Death/ReMI/MI/stroke disability, stroke, death/ReMI/MI 24 h, bleeding and ICH.	Chest discomfort within 12 h associated with transient or persistent STE or STD >0.5 mm **OR** persistent, definite T-wave inversion >1 mm.	ITT	ReMI/MI	Hirudin significantly better at 24 hrs, effect dissipated over time.

	Other Outcomes	Eligibility, inclusion criteria.	Primary analysis	Adjudicated endpoints	Results
GUSTO-3	Stroke, ICH, death/disabling stroke, death/stroke, ReMI, CHF, death 24 h and bleeding.	Onset symptoms (>30 m) within 6 h and had (based on 12-lead ECG) STE of ≥ 1 mm in ≥ 2 limb leads **OR** ≥ 2 mm in the precordial leads **OR** BBB.	ITT	Stroke	No difference.
HERO-2	Death/ReMI, ReMI, ReMI 96 h, death 24 h, stroke, bleeding and ICH.	Onset symptoms (>30 m) within 6 h, STE of ≥ 1 mm in ≥ 2 contiguous leads **OR** ≥ 2 mm in 2 contiguous precordial leads in V_1-V_3 **OR** LBBB.	ITT	ReMI, Stroke	No difference in death 30 d, significant difference in ReMI.
ASSENT-2	Death/stroke, stroke 30 d, in-hospital nonfatal cardiac events (inc. ReMI) and bleeding.	18+, onset symptoms within 6 h, STE of ≥ 0.1 mV in ≥ 2 limb leads **OR** ≥ 0.2 mV in ≥ 2 contiguous precordial leads **OR** LBBB	ITT	Stroke	Equivalence for death 30 d.
ASSENT-3	Above endpoint + ICH/major bleeding, readmission, stroke 30 d and bleeding.	18+, onset symptoms within 6 h, STE of ≥ 0.1 mV in ≥ 2 limb leads **OR** ≥ 0.2 mV in ≥ 2 contiguous precordial leads **OR** LBBB	ITT	Stroke	Significant effect for primary + major secondary.

Abbreviations: SK, Streptokinase; IV, intravenous; SC, subcutaneous; ITT, intention to treat, STE, ST-segment elevation; STD, ST-segment depression; LBBB, left bundle branch block; CHF, congestive heart failure; ICH, intracranial haemorrhage; NCB, net clinical benefit.

TABLE A2: ASSENT-3 day 0 30-day mortality model [9]

Risk factor:	χ^2	Parameter estimate (SE)
Age (years):	156	
≤65		0
65-74		0.94 (0.15)
>74		1.84 (0.15)
Race: non-Caucasian	7	0.43 (0.15)
Hypertension	15	0.45 (0.12)
Previous MI	4	0.30 (0.15)
Killip class:	32	
I		0
II		0.57 (0.15)
III-IV		1.33 (0.27)
Systolic BP (mm Hg):	75	
<120		1.14 (0.18)
120-132		0.55 (0.19)
133-150		-0.11 (0.20)
>150		0
Heart rate (bpm):	39	
≤62		0
63-73		0.29 (0.19)
74-85		0.36 (0.18)
>85		0.97 (0.17)
Anterior MI	6	0.30 (0.13)
QRS score:	33	
0-1		0
2-4		0.33 (0.21)
>4		0.52 (0.21)
ECG confounders		0.97 (0.20)
Total ST-segment deviation (mm)	16	
<12		0
12-17		0.36 (0.16)
>17		0.59 (0.16)
ECG confounders		-0.01 (0.20)

BP, blood pressure; bpm, beats per minute; ECG, electrocardiogram; MI, myocardial infarction; SE, standard error. There was a significant interaction between age and Killip class (P=0.001).

TABLE A3: *c* statistics by region for uncalibrated models

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-2	GUSTO-3
Western countries:						
HPI-FULL*	0.813	0.810	0.806	0.812	0.817	0.818
HPI-FULL- angina*†	0.813	0.810	0.806	0.813	0.816	0.819
GUSTO-1‡	0.826	0.810	0.803	0.820	0.809	0.826
ASSENT-2*	0.813	0.814	0.808	0.813	0.820	0.822
Latin America:						
HPI-FULL*	0.828	0.783	0.785			0.874
HPI-FULL- angina*†	0.827	0.783	0.788			0.867
GUSTO-1‡	0.817	0.794	0.802			0.859
ASSENT-2*	0.824	0.786	0.799			0.857
Eastern Europe:						
HPI-FULL*	0.803	0.788	0.863	0.744		0.801
HPI-FULL- angina*†	0.802	0.788	0.865	0.748		0.805
GUSTO-1‡	0.795	0.783	0.825	0.746		0.815
ASSENT-2*	0.797	0.803	0.865	0.725		0.805
Russia:						
HPI-FULL*	0.825					
HPI-FULL- angina*†	0.825					
GUSTO-1‡	0.820					
ASSENT-2*	0.822					
Asia:						
HPI-FULL*	0.766					
HPI-FULL- angina*†	0.764					
GUSTO-1‡	0.769					
ASSENT-2*	0.760					

*Treatment offset incorporated.

†HPI-FULL re-estimated without angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

TABLE A4: Overall *c* statistic results after trial-region level recalibration

	Population dataset					
Model	**HERO-2**	**ASSENT-2**	**ASSENT-3**	**GUSTO-I**	**GUSTO-2**	**GUSTO-3**
HPI-FULL*	0.818	0.809	0.807	0.811	0.817	0.818
GUSTO-1†	0.814	0.810	0.805	0.819	0.809	0.826
ASSENT-2*	0.814	0.813	0.810	0.812	0.820	0.821

*Treatment offset incorporated.

†GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

TABLE A5: Nagelkerke's R^2 statistics (%) for uncalibrated models

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-2	GUSTO-3
Overall:						
HPI-FULL*	26.9	20.6	20.4	22.6	20.1	23.6
HPI-FULL- angina*†	26.7	20.7	20.3	22.8	20.5	23.6
GUSTO-1‡	23.6	19.5	19.7	24.0	14.6	24.0
ASSENT-2*	24.4	21.3	21.2	22.7	22.4	24.0
Western patients:						
HPI-FULL*	21.2	20.2	20.0	22.7	20.1	23.5
HPI-FULL- angina*†	21.1	20.3	20.1	22.8	20.5	23.5
GUSTO-1‡	22.5	19.2	19.2	24.1	14.6	24.1
ASSENT-2*	20.8	20.8	20.7	22.8	22.4	24.0
Latin America:						
HPI-FULL*	29.5	23.0	19.2			34.5
HPI-FULL- angina*†	29.3	23.5	19.1			34.0
GUSTO-1‡	23.9	19.1	20.3			28.5
ASSENT-2*	28.4	24.4	21.6			34.0
Eastern Europe:						
HPI-FULL*	23.5	20.2	29.3	14.6		21.5
HPI-FULL- angina*†	23.3	19.4	27.9	15.1		22.3
GUSTO-1‡	20.0	20.3	24.3	14.4		21.4
ASSENT-2*	21.5	22.4	29.7	12.8		21.4
Russia:						
HPI-FULL*	29.8					
HPI-FULL- angina*†	29.7					
GUSTO-1‡	26.4					
ASSENT-2*	27.3					
Asia:						
HPI-FULL*	18.8					
HPI-FULL- angina*†	18.7					
GUSTO-1‡	10.8					
ASSENT-2*	4.0					

*Treatment offset incorporated.

†HPI-FULL re-estimated without angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

TABLE A6: Nagelkerke's R^2 statistics (%) after trial-level recalibration

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-2	GUSTO-3
Overall:						
HPI-FULL*	26.9	20.7	20.4	22.7	21.8	23.7
HPI-FULL- angina*†	26.7	20.7	20.4	22.8	21.8	23.7
GUSTO-1‡	25.1	19.8	19.9	24.0	18.6	24.3
ASSENT-2*	25.3	21.3	21.2	22.7	23.2	24.2
Western patients:						
HPI-FULL*	21.2	20.3	20.1	22.8	21.8	23.6
HPI-FULL- angina*†	21.1	20.3	20.1	22.9	21.8	23.7
GUSTO-1‡	20.6	19.5	19.4	24.1	18.6	24.2
ASSENT-2*	19.6	20.8	20.7	22.8	23.2	24.2
Latin America:						
HPI-FULL*	29.5	23.5	19.2			34.9
HPI-FULL- angina*†	29.3	23.5	19.4			34.4
GUSTO-1‡	26.9	20.5	21.2			30.2
ASSENT-2*	26.2	24.4	21.9			34.0
Eastern Europe:						
HPI-FULL*	23.5	19.8	28.8	14.2		21.3
HPI-FULL- angina*†	23.3	19.6	28.3	14.7		22.0
GUSTO-1‡	21.9	20.5	23.9	14.8		22.6
ASSENT-2*	22.5	22.4	30.0	12.1		21.7
Russia:						
HPI-FULL*	29.8					
HPI-FULL- angina*†	29.7					
GUSTO-1‡	27.9					
ASSENT-2*	28.9					
Asia:						
HPI-FULL*	18.8					
HPI-FULL- angina*†	18.7					
GUSTO-1‡	17.1					
ASSENT-2*	10.3					

*Treatment offset incorporated.

†HPI-FULL re-estimated without angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.

TABLE A7: Nagelkerke's R^2 statistics (%) after trial-region level recalibration

Model/Dataset	HERO-2	ASSENT-2	ASSENT-3	GUSTO-I	GUSTO-2	GUSTO-3
Overall:						
HPI-FULL*	26.9	20.7	20.4	22.7	21.8	23.7
HPI-FULL- angina*†	26.7	20.7	20.4	22.8	21.8	23.8
GUSTO-1‡	25.5	20.1	20.2	24.1	18.6	24.4
ASSENT-2*	26.1	21.3	21.3	22.8	23.2	24.2
Western patients:						
HPI-FULL*	21.2	20.3	20.1	22.8	21.8	23.6
HPI-FULL- angina*†	21.1	20.3	20.1	22.9	21.8	23.7
GUSTO-1‡	22.2	19.5	19.4	24.1	18.6	24.2
ASSENT-2*	20.8	20.8	20.7	22.8	23.2	24.2
Latin America:						
HPI-FULL*	29.5	23.9	19.2			35.0
HPI-FULL- angina*†	29.3	23.9	19.5			34.4
GUSTO-1‡	27.4	25.6	23.7			34.6
ASSENT-2*	28.7	24.4	22.5			34.4
Eastern Europe:						
HPI-FULL*	23.5	20.5	30.5	14.2		21.5
HPI-FULL- angina*†	23.3	20.4	30.4	14.8		22.2
GUSTO-1‡	21.9	20.5	24.4	18.0		23.4
ASSENT-2*	22.5	22.4	30.3	13.9		22.0
Russia:						
HPI-FULL*	29.8					
HPI-FULL- angina*†	29.7					
GUSTO-1‡	27.9					
ASSENT-2*	29.1					
Asia:						
HPI-FULL*	18.8					
HPI-FULL- angina*†	18.7					
GUSTO-1‡	20.0					
ASSENT-2*	18.6					

*Treatment offset incorporated before recalibration.

†HPI-FULL re-estimated without angina.

‡GUSTO-1 LP calculated ignoring history of cerebrovascular disease in ASSENT trials.