medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Development and Validation of a Multivariable Risk Prediction Model for Sudden Cardiac Death after Myocardial Infarction (PROFID Risk Model): Study Rationale, Design and Protocol

Glen P. Martin*[1], Gerhard Hindricks*[3], Artur Akbarov[1], Zoher Kapacee[1], Le Mai Parkes[1], Golnoosh Motamedi-Ghahfarokhi[1], Stephanie Ng[1], Daniel Sprague[2], Youssef Taleb[2], Marcus Ong[2], Enrico Longato[4], Christopher A. Miller[5,6,7], Alireza Sepehri Shamloo[3], Christine Albert[8], Petra Barthel[9], Serge Boveda[10,11,12], Frieder Braunschweig[13], Jens Brock Johansen[14], Nancy Cook[15], Christian de Chillou[16], Petra J.M. Elders[17], Jonas Faxen[13], Tim Friede[18,19], Laura Fusini[20], Chris P. Gale[21,22], Jiri Jarkovsky[23], Xavier Jouven[12], Juhani Junttila[24,25], Antti Kiviniemi[24], Valentina Kutyifa[26], Daniel Lee[27], Jill Leigh[28], Radosław Lenarczyk[29], Francisco Leyva[30], Michael Maeng[31], Andrea Manca[32], Eloi Marijon[33], Ursula Marschall[34], Manickavasagar Vinayagamoorthy[15], Jens Cosedis Nielsen[31], Thomas Olsen[35], Julie Pester[15], Gianluca Pontone[20], Georg Schmidt[9], Peter J. Schwartz[36], Christian Sticherling[37], Mahmoud Suleiman[38], Milos Taborsky[39], Hanno L. Tan[40,41], Jacob Tflt-Hansen[42], Jan G.P. Tijssen[43], Gordon Tomaselli[44], Tom Verstraelen[43], Kevin Kris Warnakula Olesen[31], Arthur A.M. Wilde[40], Rik Willems[45], Dick L. Willems[46], Katherine Wu[47], Markus Zabel[19,48], Niels Peek†[1], Nikolaos Dagres†[3]

* G.P.M. and G.H. contributed equally and are considered joint first authors
† N.P. and N.D. contributed equally and are considered joint senior authors

Affiliations

1. Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom
2. Spectra Analytics, London, United Kingdom
3. Department of Electrophysiology, Heart Center Leipzig at the University of Leipzig, Leipzig, Germany
4. Department of Information Engineering, University of Padova, Padova, Italy
5. Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Oxford Road, Manchester, United Kingdom
6. Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Southmoor Road, Wythenshawe, Manchester, United Kingdom
7. Wellcome Centre for Cell-Matrix Research, Division of Cell-Matrix Biology & Regenerative Medicine, School of Biology, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Oxford Road, Manchester, United Kingdom
8. Department of Cardiology, Smidt Heart Institute, Cedars Sinai Medical Center, Los Angeles, CA, USA
9. Klinikum rechts der Isar, Technische Universität München, Ismaninger Straße 22, D-81675, Munich, Germany
10. Cardiology - Heart Rhythm Management Department, Clinique Pasteur, Toulouse, France
11. Vrije Universiteit Brussel (VUB), Laarbeeklaan 101, 1090 Jette Brussels, Belgium

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

12. Paris Cardiovascular Research Center (PARCC), INSERM Unit 970, 56 Rue Leblanc, France
13. Department of Cardiology, Karolinska University Hospital, Dept of Cardiology Stockholm, Sweden
14. Department of Cardiology, Odense University Hospital, Department of Cardiology Odense, Syddanmark, Denmark
15. Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
16. Département de Cardiologie, CHRU de Nancy, Nancy F-54500, France
17. Department of General Practice and Elderly Care Medicine, Amsterdam UMC, Vrije Universiteit, Amsterdam Public Health research institute, Amsterdam, the Netherlands
18. Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany
19. German Center for Cardiovascular Research, partner site Göttingen, Göttingen, Germany
20. Department of Cardiovascular Imaging, Centro Cardiologico Monzino IRCCS, Milan, Italy
21. Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, United Kingdom
22. Department of Cardiology, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom
23. Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Czech Republic
24. Research Unit of Internal Medicine, Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland
25. Biocenter Oulu, University of Oulu, Oulu, Finland
26. University of Rochester Medical Center, Clinical Cardiovascular Research Center, Rochester, New York, USA
27. Feinberg Cardiovascular and Renal Research Institute, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
28. Boston Scientific Corporation, St. Paul, Minnesota, USA
29. Department of Cardiology, Congenital Heart Defects and Electrotherapy, Medical University of Silesia, Silesian Center of Heart Disease, Zabrze, Poland
30. Aston Medical School, Aston University, Aston Triangle, Birmingham, United Kingdom
31. Department of Clinical Medicine and Department of Cardiology, Aarhus University, Aarhus, Denmark
32. Centre for Health Economics, University of York, York, United Kingdom
33. European Georges Pompidou Hospital and University of Paris, Paris, France
34. Department of Medicine and Health Services Research, BARMER Health Insurance, Lichtscheider Strasse 89, 42285 Wuppertal, Germany
35. Department of Cardiology, Odense University Hospital, Odense, Denmark
36. Istituto Auxologico Italiano, IRCCS, Center for Cardiac Arrhythmias of Genetic Origin, Milan, Italy
37. University Hospital Basel, University Basel
38. Department of Cardiology, Rambam Health Care Campus, Haifa, Israel
39. Department of Internal Medicine I – Cardiology, Olomouc University Hospital, Olomouc, Czech Republic

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

40. Dept of Clinical and Experimental Cardiology, Amsterdam University Medical Center location AMC, Amsterdam, the Netherlands
41. Netherlands Heart Institute, Utrecht, the Netherlands
42. The Department of Cardiology, The Heart Centre, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
43. Department of Cardiology, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
44. The Albert Einstein College of Medicine, Bronx, New York, United States
45. University Hospitals (UZ) Leuven, Leuven
46. Department of Ethics, Law and Humanities, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, the Netherlands
47. Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
48. Department of Cardiology and Pneumology, Heart Center, University Medical Center Goettingen, Robert-Koch-Strasse 40, 37075 Goettingen, Germany

Contributorship statement

Gerhard Hindricks and Nikolaos Dagres conceived the overall idea of the PROFID project and are the study principal investigators. Glen P. Martin, Zoher Kapacee and Niels Peek drafted the initial version of this manuscript, with scientific input on the study design and methodology from Artur Akbarov, Le Mai Parkes, Golnoosh Motmedi-Ghahfarokhi, Stephanie Ng, Daniel Sprague, Youssef Taleb, Marcus Ong, Enrico Longato and Chris Miller. All authors critically reviewed the manuscript for scientific content and have approved the final version.

Competing Interests

The authors have no conflicts of interest to declare.

Funding

The work described is part of the PROFID project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 847999.

Word Count

3898

Corresponding Author

Dr Glen Philip Martin
Senior Lecturer in Health Data Science
Vaughan House, University of Manchester, Manchester, M13 9GB, United Kingdom
Email: glen.martin@manchester.ac.uk

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

Introduction

Sudden cardiac death (SCD) is the leading cause of death in patients with myocardial infarction (MI) and can be prevented by the implantable cardioverter defibrillator (ICD). Currently, risk stratification for SCD and decision on ICD implantation are based solely on impaired left ventricular ejection fraction (LVEF). However, this strategy leads to over- and under-treatment of patients because LVEF alone is insufficient for accurate assessment of prognosis. Thus, there is a need for better risk stratification. This is the study protocol for developing and validating a prediction model for risk of SCD in patients with prior MI.

Methods and Analysis

The EU funded PROFID project will analyse 23 datasets from Europe, Israel and the US (~225,000 observations). The datasets include patients with prior MI or ischemic cardiomyopathy with reduced LVEF<50%, with and without a primary prevention ICD. Our primary outcome is SCD in patients without an ICD, or appropriate ICD therapy in patients carrying an ICD as a SCD surrogate. For analysis, we will stack 18 of the datasets into a single database (datastack), with the remaining analysed remotely for data governance reasons (remote data). We will apply 5 analytical approaches to develop the risk prediction model in the datastack and the remote datasets, all under a competing risk framework: 1) Weibull model, 2) flexible parametric survival model, 3) random forest, 4) likelihood boosting machine, and 5) neural network. These dataset-specific models will be combined into a single model (one per analysis method) using model aggregation methods, which will be externally validated using systematic leave-one-dataset-out cross-validation. Predictive performance will be pooled using random effects meta-analysis to select the model with best performance.

Ethics and dissemination

Local ethical approval was obtained. The final model will be disseminated through scientific publications and a web-calculator. Statistical code will be published through open-source repositories.

Keywords

Clinical prediction model; model development and validation; sudden cardiac death; myocardial infarction; protocol, defibrillator implantation

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Introduction

Sudden cardiac death (SCD) is the leading cause of death,[1] accounting for approximately 20% of all deaths in Europe [2,3], and the estimated yearly incidence of SCD in European countries is around 1 per 1,000 inhabitants,[4] afflicting approximately 350,000-700,000 Europeans annually.[3–6] The majority of SCD cases occur in people with coronary artery disease and are mostly caused by ventricular tachyarrhythmias following myocardial infarction (MI). After MI, a reduced left ventricular ejection fraction (LVEF) is associated with increased risk for all-cause death, cardiac death and SCD.[8] Randomised clinical trials (RCTs) have demonstrated that in patients with severely impaired LVEF, the risk of SCD and all-cause death may be significantly reduced through prophylactic implantation of an implantable cardioverter defibrillator (ICD) [9–11]. Based on these data, international guidelines recommend ICD implantation in post-MI patients with severely reduced LVEF ≤35% for primary prevention of SCD.[1]

However, since completion of these landmark trials in the late 1990s and early 2000's, major advances in pharmacological and non-pharmacological treatment have led to substantial decline in the risk of SCD after MI.[1,3] Concordantly, only a minority of these patients will ever need the implanted ICD and experience appropriate ICD therapies. In fact, the numerical majority of SCD cases occur in patients with LVEF>35% and who are not considered for a primary prevention ICD implantation according to current guidelines [3,12]. Thus, there is consensus that the current practice of using LVEF as the sole risk stratification factor for the risk of SCD after MI and the decision on prophylactic ICD implantation has significant limitations. [13]

Other clinical characteristics, laboratory and imaging biomarkers, genetic markers and risk factors have been reported to be associated with increased risk of SCD.[3,13,14] However, in isolation, none of these prognostic factors have sufficient predictive accuracy for clinical use. Consequently, there is an imperative and unmet need for the development of a clinical prediction model (CPM) to use a combination of such predictor variables to estimate the risk of SCD after MI. Existing CPMs [15–19] have not been implemented into recommended practice, largely because improvements in their predictive ability (and clinical utility [20]) is required.

The development of a CPM for risk of SCD in patients with previous MI is a primary objective of the PROFID project – a large Horizon 2020 funded pan-European consortium [21] (grant no. 847999). The aim of this CPM is to aid the risk stratification of patients with MI and facilitate decision-making for primary prevention ICD implantation. Once the PROFID CPM has been developed, the PROFID project will compare its predictive performance with the existing CPMs (including using LVEF as the sole risk stratification variable, reflecting current clinical practice). The project will then compare personalised decision-making for prophylactic ICD implantation with application of the CPM against current clinical practice in patients with LVEF≤35% and LVEF>35% in two multinational randomised clinical trials, (PROFID-Reduced and PROFID-Preserved, NCT04540354 and NCT04540289).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

The aim of this paper is to describe the study protocol for the development and validation of the PROFID CPM, including the key analytical steps and decisions that have been considered.

Methods and Analysis

We structure this protocol in line with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement and checklist.[22] Given that this is a protocol, we focus on the methods section of the TRIPOD statement. Subsequent publications reporting the results will adhere to the full TRIPOD checklist.

Aim

The aim of the study, on which this protocol is based, is to develop and validate a multivariable CPM to estimate the risk of SCD after MI, using data from Europe, Israel and the US.

Study Design and Data Sources

This will be a retrospective analysis of data arising from observational cohort studies, routine healthcare records and randomised controlled trials. **Table 1** shows the full list of datasets used for the development of the CPM, comprising 23 datasets from 12 countries. All datasets describe individuals who have had an MI or have coronary artery disease with ischemic cardiomyopathy and reduced LVEF <50%, and all contain information on SCD (or surrogates thereof). We distinguish between four types of datasets: (i) acute MI cohort datasets, where each patient was entered into the dataset at time of their acute MI event; (ii) prior MI or ischemic cardiomyopathy cohort datasets, where patient had a previous MI or ischemic cardiomyopathy and reduced left ventricular ejection fraction (<50% as defined by current heart failure guidelines of the European Society of Cardiology [23]) and were entered into the dataset at some time after their MI event; (iii) ICD cohort datasets, where each patient was entered into the dataset at the time of receiving their prophylactic ICD implant for primary prevention of SCD after an MI; and (iv) datasets from randomised controlled trials in which participants with prior MI or ischemic cardiomyopathy and severely reduced LVEF, received ICD implants or medication therapy (e.g. MADIT II,[9,10], SCD-HeFT [11]). All datasets contain clinical information, medications and other selected variables such as ECG parameters and biomarkers, which were recorded for each patient at time of entry into the dataset. Some datasets include additional variables regarding cardiac magnetic resonance (CMR) imaging, which have been shown to be highly prognostic of SCD [24] (see **Table 1**). All data include information about outcomes during follow-up for each patient: until time of SCD, death from other causes, or until time of first appropriate ICD shocks/anti-tachycardia pacing, hereafter referred to as appropriate therapy.

Prior to analysis, all datasets will undergo data cleaning in partnership and agreement with the respective data providers. Given that each of the datasets contain different variables, we have developed a common data model to ensure that there is a consistent set of variables across all data sources (**Supplementary Table 1**). In particular, the common data model

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

dictates the units of measurement where applicable, categories for nominal and ordinal variables, and definitions of each variable. The variables listed in the common data model define the list of candidate predictor variables we will consider for inclusion in the model (i.e., for variable selection – see supplementary methods). We will also include calendar year in the models as a continuous variable to account for varying risk through time. At the time of prediction, this will be always set to the latest calendar year in the development data.

Participant Entry Criteria

Our analysis will include all patients who are at least 18 years old and have had either: (a) a previous MI defined as ST-segment-elevation myocardial infarction or non-ST-segment-elevation myocardial infarction, or (b) coronary artery disease and ischemic cardiomyopathy with reduced left ventricular ejection fraction <50% (as defined by current heart failure guidelines of the European Society of Cardiology [23]). From these patients, we exclude: (i) patients who received ICD implantation for secondary prevention of SCD at baseline, (ii) patients with a cardiac resynchronization therapy (CRT) device at baseline, (iii) patients with non-ischemic cardiomyopathy such as dilated cardiomyopathy, hypertrophic cardiomyopathy, or restrictive cardiomyopathy, (iv) patients with a primary electrical arrhythmic disease such as long QT syndrome, Brugada syndrome or catecholaminergic polymorphic VT, (v) patients with congenital heart disease, and (vi) patients who died (or experience the outcome) within the first 40 days after the index MI. A schematic of the inclusion and exclusion criteria is given in **Figure 1**.

Patients with a CRT device at baseline are excluded because CRT reduces the risk for SCD [25]. The reason for exclusion criterion (vi) is that, after MI, patients undergo a ventricular remodelling period, during which important parameters for the subsequent risk such as LVEF may significantly change. Therefore, decisions on prophylactic ICD implantation are made after ventricular remodelling. Since the intention is to use the PROFID CPM after ventricular remodelling, we designed the study to match this prediction time point (i.e., predictions are possible any time after the 40-day remodelling period).

Outcome definitions and analytic approach

Our primary outcome is time-to-sudden cardiac death or to life-threatening ventricular arrhythmias (ventricular tachycardia, or ventricular fibrillation) during follow-up. Time-zero for calculating the time-to-event outcomes will be defined as follows:

1. In the acute MI cohort datasets, time-zero for each patient is 40 days after the index MI event. If an individual has more than one MI event within the dataset, then their index MI will be chosen randomly across all their MI events prior to their first competing risk outcome. The decision to take a random MI event ensures maximum consistency across datasets and avoids introducing biases if one were to select the first or last MI.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

2. In the prior MI or ischaemic cardiomyopathy cohort datasets, time-zero for each patient is their time of entry into the dataset/study, provided it is more than 40 days after the initial MI event.

3. In ICD datasets, time-zero for each patient is the time of ICD implantation for primary prevention of SCD, provided it is more than 40 days after the initial MI event.

The primary outcome for each patient is then defined as the time between time-zero and SCD or life-threatening ventricular arrhythmia event, which is defined across our datasets as follows, depending on whether a patient has an ICD implant or not:

- In patients without an ICD, the primary outcome is occurrence of SCD based on cause-of-death adjudication, or implantation of ICD for secondary prevention of SCD (i.e., due to the occurrence of ventricular tachycardia or ventricular fibrillation). The definitions for which cause-of-deaths were listed as SCD varies across datasets. Moreover, the Swedish Heart Registry does not contain information on SCD directly, rather whether the patient had a successful resuscitation outcome for sudden cardiac arrest, which will be used as a surrogate for SCD (i.e., the first resuscitation event will be used as the endpoint for that individual). We therefore expect to see heterogeneity in the incidence of SCD across the included datasets, which will be accounted for directly through the modelling (see below).
- In patients with an ICD, the primary outcome will be defined as appropriate therapy delivered by an ICD. This surrogate for SCD is required since patients with an ICD cannot experience SCD that is preventable by the device. First appropriate therapy mainly includes first appropriate shock and/or the first appropriate anti-tachycardia pacing (ATP); however, some datasets do not contain information on ATP, where first appropriate therapy will include appropriate shock only. We will account for such differences within the models (see below). A limitation of using appropriate therapy is that the programming of each ICD device differs across datasets and patients. However, appropriate ICD therapies remain the best available surrogate for SCD in patients with an ICD, particularly considering the objective of applying the CPM for decision-making on need for ICD implantation.

For any patient who receives an ICD for primary prevention of SCD during their follow-up (i.e., without prior occurrence of ventricular tachycardia or ventricular fibrillation), the data on appropriate therapies after the time of primary prevention ICD implant will be used to define their time-to-event outcome. If a patient receives an ICD during follow-up and we do not have data on subsequent appropriate therapies, then we will censor this patient at the time of ICD implantation, unless this ICD was implanted for secondary prevention (ventricular tachycardia or ventricular fibrillation) in which case it will be taken as our primary outcome; such censoring will be only required in the Swedish Heart Registry.

In all time-to-event analyses, we will account for competing risks of death from other causes using the Fine and Gray competing risks modelling framework.[26] Specifically, in patients without an ICD, we distinguish (i) SCD (resuscitated or non-resuscitated, including sustained

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

VT/VF), and (ii) death from any other cause, where (ii) is a competing risk for (i). In patients with an ICD, we distinguish (i) appropriate therapy (shock – with or without ATP), and (ii) death from any other cause including SCD that was not prevented by an ICD.

Heart transplantation or implantation of left ventricular mechanical assist device during follow-up will be considered a censoring event. All other censoring will be administrative censoring.

The differences in the outcome definition across patients with and without ICDs causes heterogeneity and clustering in our analysis, which we will account for during the modelling by including the following patient-level binary indicators as covariates: (i) entry of study with an ICD vs. entry of study without an ICD, (ii) patient from the Swedish Heart Registry (resuscitation outcomes only) vs. patient not from the Swedish Heart Registry, and (iii) appropriate therapy defined as shock only vs. shock or anti-tachycardia pacing. We will also include interactions in the model between LVEF and the timing of such measurement (before or after 40-day ventricular remodelling period).

Sample Size calculation

Sample size criteria for the development of CPMs for continuous, binary and time-to-event outcomes have recently been proposed.[27–29]. Since sample size criteria for competing risk models (and non-regression-based models) have yet to be developed, we base our sample size calculation on the criteria developed for time-to-event prediction models.[28] We assumed the following when making our sample size calculations for developing a time-to-event CPM (some values, such as event proportions, are based on the datasets available to us): (i) 2.2% experience SCD during follow-up, (ii) the mean follow-up time is 4.5 years, (iii) the model would explain 15% of maximum R^2, (iv) we target a maximum degree of shrinkage/overfitting of 0.9, and (v) we have approximately 100 candidate predictors in our common data model. This resulted in a minimum required sample size of 24219, which was driven by criteria 1 of Riley et al.[28] Combined, the PROFID datasets include approximately 225,000 patients, which far exceeds the minimum requirements.

Missing Data

Many of the datasets in this study will contain variables with missing values. During the exploratory and descriptive analyses, we will investigate missing data patterns using graphical plots and tabulations. During CPM development, validation, and deployment, we will use fuzzy K-means to impute missing data,[30] which has been shown to be a robust method in prediction context.[31] For this study, fuzzy K-means is especially attractive over alternative imputation methods because it can be easily implemented during model deployment since it only requires cluster centroids to be retained (which poses no risk in terms of information governance). While multiple imputation would be an alternative imputation strategy for developing the CPM, it poses issues when deploying the model in practice, because to use the CPM with missing data would require one to store a copy of the development data, and as such will not be considered here. [32] Overall, the data contains both *sporadically missing values*, (variables with missing values for some patients within a

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

dataset) and *systematically missing values* (variables that are completely missing for a whole dataset, e.g., where the datasets did not record a particular variable). For *sporadically missing* values, we will assume they are missing at random. We will assume *systematically missing values* are missing completely at random[33].

Statistical Analysis Methods

All our analysis choices have required us to respect that we will not have direct access to the individual participant data (IPD) from all datasets; we will only be able to access some datasets through remote analysis whereby our analytical scripts are sent to the data custodians, who run them and return to us the analytical results. Specifically, we have direct (IPD-level) access to 18 of the 23 datasets, which we will "stack" into a single database (hereafter named 'datastack'), while the remaining 5 datasets will be analysed remotely for data governance reasons ('remote data'). To account for this approach to data access, we have developed a bespoke two-phase design, following best-practice recommendations.[34,35] We provide an overview of the statistical processes in this section. A graphical representation of our modelling approach is given in **Figure 2**.

In all assessments of predictive performance that we mention below, we quantify this using discrimination, calibration and overall accuracy; all measures will be estimated within the competing risk framework.[36] We will quantify discrimination of the prediction model using the weighted Harrell's weighted C-index.[37] We will assess the calibration using calibration plots of the observed survival curves (Kaplan-Meier plot) against those predicted by the model. Where relevant (i.e. for regression-based methods), we will also summarise calibration by estimating the calibration slope (ideal value 1) by fitting a Cox regression model (to the sub-distribution hazard of SCD) to the observed outcomes and with the linear predictor (for regression models) from the model as the only covariate. Finally, we will compute overall accuracy through the integrated Brier score. The inverse probability of censoring weights versions of both Harrell's C-index and the integrated Brier score will be based on Kaplan-Meier estimates of the censoring distribution in the training sets.

Phase 0: data preparation and ratification

From the onset of our analysis, within all datasets we will temporally hold-out 10% of data using an event-stratified sampling approach based on the latest event times (i.e., where we randomly select the latest 10% of the temporally ordered event dates), which is similar to period analysis [38]. Hereafter, these are all called the '10% hold-out sets', with the remaining 90% of each dataset called the 'development sets' (**Figure 2**). The 10% hold-out sets will be used in the second phase of modelling to temporally validate the selected analytical model from phase 1 (see below). We emphasise that within each modelling phase our approach to validating CPMs is systematic internal-external cross-validation.[39,40] As a preliminary modelling step, we will undertake a data ratification exercise, where reports will be sent to each data provider to ensure consistency in approaches to cleaning and analyzing each data source.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Phase 1: model development, aggregation, and systematic internal-external cross-validation

For this first phase of the main analysis, we use model aggregation methods combined with systematic inter-external cross-validation (across the 90% development subsets) [39,40] to select the "best" analytical method to take forward for the final PROFID CPM (**Figure 2**). Specifically, this process involves leaving out one of the 90% development subsets, with the remaining development subsets used to develop a CPM on (a) the datastack combined and (b) on each remote dataset in turn, using each of the following methods (within a competing risk framework): 1) Weibull modelling, 2) flexible parametric survival modelling [41,42], 3) random survival forests [43], 4) likelihood boosting machine, and 5) neural network [44] (see the supplementary methods for details on how we will fit each method and variable selection). For each of the five analytical methods, this will create either 5 or 6 CPMs in each iteration of the leave-one-out-cross-validation (5 if the left-out set is a remote dataset, and 6 if the left-out set is within the datastack, since we have the datastack plus 5 other remote datasets). Since we need a single PROFID model, these 5/6 CPMs will then be combined using the methods described by Debray et al [45] (see supplementary methods) to create five aggregated CPMs (one aggregated CPM per analytical method). Each of these five aggregate models will then be validated in the left-out 90% development subset. This process then cycles through leaving out each development subset in-turn, resulting in 23 sets of predictive performance estimates per analytical method. We will use random effects meta-analysis to combine these estimates, resulting in 5 sets of pooled (meta-anlaysed) predictive performance estimates, with associated measures of heterogeneity (I^2) and prediction intervals (i.e. the potential model performance in a new population similar to those included in the meta□analysis). [46]

At this point, we will consider leaving out some datasets if the hetrogenity assessment indicates some data are markedly different (detrimentally) to the others. Any such decisions will also be based on clinical assessments, to avoid this being a completely data-driven exercise. If we do leave out a dataset, then the aforementioned internal-external validation processes would be repeated.

After completing this step, we will use the pooled (meta-analysed) predictive performance measures to select the analytical method for the final PROFID CPM. We will favour the analytical method that leads to the highest predictive performance in terms of calibration and discrimination. If multiple methods lead to similar performance or if different performance metrics favour different models, then we will favour the model that has the fewest number of predictor variables and is most transparent in how it calculates risk estimates (interpretability) and has the easiest clinical implementation.

Phase 2: model updating for CMR variables and final model systematic internal-external cross-validation

At the end of phase 1, we will have selected the analytical method that will be used for the final PROFID CPM. In phase 2, we will seek to use the temporally held-out sets (10% hold-out sets) to obtain unbiased estimates of predictive performance of said model (having used the prediction performance estimates from phase 1 to select an analytical method).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Specifically, we will take the selected analytical method (and corresponding models) from phase 1 and calculate the performance in each of the 10% temporal hold-out sets separately; to estimate genuine temporal performance, we will not use the 90% development subsets further here, but rather take the models as developed from phase 1. As in phase 1, we will apply meta-analysis to pool the estimates of predictive performance of the final aggregate model [46]; this pooled set of performance estimates become our final external validation measures, and we will also calculate prediction intervals for each performance estimate (i.e. the potential model performance in a new population similar to those included in the meta-analysis [47]).

Additionally, phase 2 will also consider the addition of the CMR variables into the modelling (using six datasets that have such data recorded). Having determined the best analytical approach (phase 1), we will again employ the systematic internal-external cross-validation framework on the subset of datasets with CMR variables to perform flexible parametric extension and recalibration [48–54] thereby allowing inclusion of CMR variables into the model. Specifically, we will train a flexible parametric survival model with the following covariates: the logit-transformed 1-year probability output of the selected phase 1 model, core scar estimated with the full-width half-maximum (FWHM) method, and grayzone. Maintaining appropriate data segregation to avoid information leakage, we will impute core scar in the one dataset without FWHM (Centro Cardiologico Monzino Registry), whereas we will use quantile normalisation to obtain a unified representation of grayzone across all datasets, where several different methods were used for quantification. Iterating across all leave-one-out datasets, we will then apply meta-analysis to pool predictive performance of this extended model.

Finally, using the 10% temporal hold-out sets, we will compare the 1-year predicted risks for individuals from the CPM with CMR data and the CPM without CMR data, thus providing the likelihood that the 2.5% risk threshold (for the PROFID-Reduced trial) or 3% threshold (for the PROFID-Preserved trial) be crossed after the addition of CMR given the initial PROFID CPM prediction. This will inform the decision of subgroups of patients (combinations of risk variables) where requesting CMR data at prediction time would be informative.

As a final analytical step, we will use 100% of the datastack and of each remote dataset (again combining the models across these using model aggregation) to fit the final PROFID, using the best analytical method.

Model Output

The output of the PROFID model will be i) risk of SCD preventable by an ICD implant, corrected for risk of death by competing causes, and ii) risk of death by competing causes, corrected for risk of SCD. For context, we will also report overall mortality risk (i.e., risk of SCD + risk of death from other causes).

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Ethics and Dissemination

All patient data analysed in this study will be de-identified and stored in a secure data environment for analysis; all individual participating databases were approved by their respective ethical boards, whenever required by national or local regulations strictly in adherence to GDPR, European laws and local data safety protocols.

We will disseminate our results through scientific publications. All statistical code for the analysis will be made available through open-source repositories (e.g., git/github) for full transparency and reproducibility. We will not be able to share the individual participant data owing to the required data sharing agreements and considerations. Finally, we will embed the final model in an appropriate open-source platform (e.g., web-calculator/ app). All dissemination will follow the TRIPOD guidelines.[22]

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

1 Priori SG, Blomström-Lundqvist C, Mazzanti A, *et al.* 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the Europe. *Europace* 2015;**17**. doi:10.1093/europace/euv319

2 Gorgels APM, Gijsbers C, De Vreede-Swagemakers J, *et al.* Out-of-hospital cardiac arrest - The relevance of heart failure. The Maastricht Circulatory Arrest Registry. *Eur Heart J* 2003;**24**. doi:10.1016/S0195-668X(03)00191-X

3 Wellens HJJ, Schwartz PJ, Lindemans FW, *et al.* Risk stratification for sudden cardiac death: Current status and challenges for the future. *Eur Heart J* 2014;**35**. doi:10.1093/eurheartj/ehu176

4 De Vreede-Swagemakers JJM, Gorgels APM, Dubois-Arbouw WI, *et al.* Out-of-Hospital Cardiac Arrest in the 1990s: A Population-Based Study in the Maastricht Area on Incidence, Characteristics and Survival. 1997. www.acc.org/members

5 Graf J, Mühlhoff C, Doig GS, *et al.* Health care costs, long-term survival, and quality of life following intensive care unit admission after cardiac arrest. *Crit Care* 2008;**12**:1–9. doi:10.1186/cc6963

6 Berdowski J, Berg RA, Tijssen JGP, *et al.* Global incidences of out-of-hospital cardiac arrest and survival rates: Systematic review of 67 prospective studies. *Resuscitation* 2010;**81**. doi:10.1016/j.resuscitation.2010.08.006

7 Zipes DP and WHJ. *Sudden cardiac death.* 2000.

8 Bauer A, Barthel P, Schneider R, *et al.* Improved Stratification of Autonomic Regulation for risk prediction in post-infarction patients with preserved left ventricular function (ISAR-Risk). *Eur Heart J* 2009;**30**:576–83. doi:10.1093/eurheartj/ehn540

9 Moss AJ, Cannom DS, Daubert JP, *et al.* Multicenter automatic defibrillator implantation trial II (MADIT II): Design and clinical protocol. *Ann Noninvasive Electrocardiol* 1999;**4**:83–91. doi:10.1111/j.1542-474X.1999.tb00369.x

10 Goldenberg I, Vyas AK, Hall WJ, *et al.* Risk Stratification for Primary Implantation of a Cardioverter-Defibrillator in Patients With Ischemic Left Ventricular Dysfunction. *J Am Coll Cardiol* 2008;**51**:288–96. doi:10.1016/j.jacc.2007.08.058

11 Bardy GH, Lee KL, Mark DB, *et al.* Amiodarone or an Implantable Cardioverter-Defibrillator for Congestive Heart Failure. 2005. www.nejm.org

12 Fishman GI, Chugh SS, Dimarco JP, *et al.* Sudden cardiac death prediction and prevention: Report from a national heart, lung, and blood institute and heart rhythm society workshop. *Circulation* 2010;**122**. doi:10.1161/CIRCULATIONAHA.110.976092

13 Dagres N, Hindricks G. Risk stratification after myocardial infarction: Is left ventricular ejection fraction enough to prevent sudden cardiac death? Eur. Heart J. 2013;**34**. doi:10.1093/eurheartj/eht109

14 Goldberger JJ, Buxton AE, Cain M, *et al.* Risk stratification for arrhythmic sudden cardiac death: Identifying the roadblocks. *Circulation* 2011;**123**. doi:10.1161/CIRCULATIONAHA.110.959734

15 Bilchick KC, Wang Y, Cheng A, *et al.* Seattle Heart Failure and Proportional Risk Models Predict Benefit From Implantable Cardioverter-Defibrillators. *J Am Coll Cardiol* 2017;**69**:2606–18.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

doi:10.1016/j.jacc.2017.03.568

16 Van Rees JB, Borleffs CJW, Van Welsenes GH, *et al.* Clinical prediction model for death prior to appropriate therapy in primary prevention implantable cardioverter defibrillator patients with ischaemic heart disease: The FADES risk score. *Heart* 2012;**98**:872–7. doi:10.1136/heartjnl-2011-300632

17 S. LD, Judy H, Raymond Y, *et al.* Clinical Risk Stratification for Primary Prevention Implantable Cardioverter Defibrillators. *Circ Hear Fail* 2015;**8**:927–37. doi:10.1161/CIRCHEARTFAILURE.115.002414

18 Verstraelen TE, van Barreveld M, van Dessel PHFM, *et al.* Development and external validation of prediction models to predict implantable cardioverter-defibrillator efficacy in primary prevention of sudden cardiac death. *EP Eur* Published Online First: 2021. doi:10.1093/europace/euab012

19 Barsheshet A, Moss AJ, Huang DT, *et al.* Applicability of a Risk Score for Prediction of the Long-Term (8-Year) Benefit of the Implantable Cardioverter-Defibrillator. *J Am Coll Cardiol* 2012;**59**:2075–9. doi:https://doi.org/10.1016/j.jacc.2012.02.036

20 Sachs MC, Sjölander A, Gabriel EE. Aim for Clinical Utility, Not Just Predictive Accuracy. *Epidemiology* 2020;**31**.https://journals.lww.com/epidem/Fulltext/2020/05000/Aim_for_Clinical_Utility,_Not_Just_Predictive.8.aspx

21 Dagres N, Peek N, Leclercq C, *et al.* The PROFID project. *Eur Heart J* 2020;**41**:3781–2. doi:10.1093/eurheartj/ehaa645

22 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol* 2015;**67**. doi:10.1016/j.eururo.2014.11.025

23 Ponikowski P, Voors AA, Anker SD, *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur. Heart J. 2016;**37**:2129-2200m. doi:10.1093/eurheartj/ehw128

24 Zegard A, Okafor O, de Bono J, *et al.* Myocardial Fibrosis as a Predictor of Sudden Death in Patients With Coronary Artery Disease. *J Am Coll Cardiol* 2021;**77**:29–41. doi:10.1016/j.jacc.2020.10.046

25 Cleland JGF, Daubert J-C, Erdmann E, *et al.* Longer-term effects of cardiac resynchronization therapy on mortality in heart failure [the CArdiac REsynchronization-Heart Failure (CARE-HF) trial extension phase]. *Eur Heart J* 2006;**27**:1928–32. doi:10.1093/eurheartj/ehl099

26 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 1999;**94**:496–509. doi:10.1080/01621459.1999.10474144

27 Riley RD, Snell KIE, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Stat Med* 2019;**38**. doi:10.1002/sim.7993

28 Riley RD, Snell KIE, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;**38**:1276–96. doi:10.1002/sim.7992

29 Riley RD, Ensor J, Snell KIE, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;**368**. doi:10.1136/bmj.m441

30 Li D, Deogun J, Spaulding W, *et al.* Towards missing data imputation: A study of fuzzy K-means

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

clustering method. *Lect Notes Artif Intell (Subseries Lect Notes Comput Sci* 2004;**3066**:573–9. doi:10.1007/978-3-540-25929-9_70

31 Mandel J SP. A Comparison of Six Methods for Missing Data Imputation. *J Biom Biostat* 2015;**06**:1–6. doi:10.4172/2155-6180.1000224

32 Sperrin M, Martin GP, Sisk R, *et al.* Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol* 2020;**125**:183–7. doi:10.1016/j.jclinepi.2020.03.028

33 Rubin DB. Biometrika Trust Inference and Missing Data Author (s): Donald B . Rubin Published by: Oxford University Press on behalf of Biometrika Trust Stable URL: http://www.jstor.org/stable/2335739 Accessed: 12-06-2016 21: 34 UTC. *Biometrika* 1976;**63**:581–92. doi:10.1186/1471-2105-12-432

34 Steyerberg EW. *Clinical Prediction Models*. NY: : Springer New York 2009. doi:10.1007/978-0-387-77244-8

35 Riley RD, Windt D, Croft P, *et al. Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford University Press 2019.

36 Wolbers M, Blanche P, Koller MT, *et al.* Concordance for prognostic models with competing risks. *Biostatistics* 2014;**15**:526–39. doi:10.1093/biostatistics/kxt059

37 Harrell FE, Califf RM, Pryor DB, *et al.* Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc* 1982;**247**:2543–6. doi:10.1001/jama.1982.03320430047030

38 Brenner HGO. Deriving More Up-to-Date Estimates of Long-Term Patient Survival. 1997.

39 Takada T, Nijman S, Denaxas S, *et al.* Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *J Clin Epidemiol* Published Online First: April 2021. doi:10.1016/j.jclinepi.2021.03.025

40 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J. Clin. Epidemiol. 2016;**69**. doi:10.1016/j.jclinepi.2015.04.005

41 Nelson CP, Lambert PC, Squire IB, *et al.* Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med* 2007;**26**:5486–98. doi:10.1002/sim.3064

42 Mozumder SI, Rutherford MJ, Lambert PC. stpm2cr: A flexible parametric competing risks model using a direct likelihood approach for the cause-specific cumulative incidence function. 2017.

43 Ishwaran H, Kogalur UB, Blackstone EH, *et al.* Random survival forests. *Ann Appl Stat* 2008;**2**:841–60. doi:10.1214/08-AOAS169

44 Lee C, Zame WR, Yoon J, *et al.* DeepHit: A deep learning approach to survival analysis with competing risks. In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. 2018.

45 Debray TPA, Moons KGM, Ahmed I, *et al.* A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;**32**:3158–80. doi:10.1002/sim.5732

46 Snell KIE, Ensor J, Debray TPA, *et al.* Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2018;**27**:3505–22. doi:10.1177/0962280217705678

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

47 Ensor J, Snell KIE, Debray TPA, *et al.* Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Stat Med* 2021;**40**:3066–84. doi:https://doi.org/10.1002/sim.8959

48 Nieboer D, Vergouwe Y, Ankerst DP, *et al.* Improving prediction models with new markers: A comparison of updating strategies. *BMC Med Res Methodol* 2016;**16**:1–10. doi:10.1186/s12874-016-0231-2

49 Grant SW, Hickey GL, Head SJ. Statistical primer: Multivariable regression considerations and pitfalls. *Eur J Cardio-thoracic Surg* 2019;**55**:179–85. doi:10.1093/ejcts/ezy403

50 Janssen KJM, Moons KGM, Kalkman CJ, *et al.* Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;**61**:76–86. doi:10.1016/j.jclinepi.2007.04.018

51 Moons KGM, Kengne AP, Grobbee DE, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;**98**:691–8. doi:10.1136/heartjnl-2011-301247

52 Van Calster B, Van Hoorde K, Vergouwe Y, *et al.* Validation and updating of risk models based on multinomial logistic regression. *Diagnostic Progn Res* 2017;**1**:1–14. doi:10.1186/s41512-016-0002-x

53 Sim J, Teece L, Dennis MS, *et al.* Validation and recalibration of two multivariable prognostic models for survival and independence in acute stroke. *PLoS One* 2016;**11**:1–17. doi:10.1371/journal.pone.0153527

54 Van Kempen BJH, S. Ferket B, Kavousi M, *et al.* Performance of Framingham cardiovascular disease (CVD) predictions in the Rotterdam Study taking into account competing risks and disentangling CVD into coronary heart disease (CHD) and stroke. *Int J Cardiol* 2014;**171**:413–8. doi:10.1016/j.ijcard.2013.12.036

55 Junttila MJ, Kiviniemi AM, Lepojärvi ES, *et al.* Type 2 diabetes and coronary artery disease: Preserved ejection fraction and sudden cardiac death. *Hear Rhythm* 2018;**15**:1450–6. doi:10.1016/j.hrthm.2018.06.017

56 Kiviniemi AM, Lepojärvi ES, Tulppo MP, *et al.* Prediabetes and Risk for Cardiac Death Among Patients With Coronary Artery Disease: The ARTEMIS Study. Published Online First: 2019. doi:10.2337/dc18-2549

57 Sticherling C, Arendacka B, Svendsen JH, *et al.* Sex differences in outcomes of primary prevention implantable cardioverter-defibrillator therapy: combined registry data from eleven European countries. doi:10.1093/europace/eux176

58 Boveda S, Narayanan K, Jacob S, *et al.* Temporal Trends Over a Decade of Defibrillator Therapy for Primary Prevention in Community Practice. *J Cardiovasc Electrophysiol* 2017;**28**:666–73. doi:10.1111/jce.13198

59 Sabbag A, Glikson M, Suleiman M, *et al.* Arrhythmic burden among asymptomatic patients with ischemic cardiomyopathy and an implantable cardioverter-defibrillator. *Hear Rhythm* 2019;**16**:813–9. doi:10.1016/j.hrthm.2019.03.030

60 Moss AJ, Zareba W, Jackson Hall W, *et al.* Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *N Engl J Med* 2002;**346**:877–83. doi:10.1056/NEJMoa013474

61 Codreanu A, Odille F, Aliot E, *et al.* Electroanatomic Characterization of Post-Infarct Scars.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Comparison With 3-Dimensional Myocardial Scar Reconstruction Based on Magnetic Resonance Imaging. *J Am Coll Cardiol* 2008;**52**:839–42. doi:10.1016/j.jacc.2008.05.038

62 Hinkle LE, Tzvi Thaler H. Clinical Classification of Cardiac Deaths. http://ahajournals.org

63 Chatterjee NA, Moorthy MV, Pester J, *et al.* Sudden death in patients with coronary heart disease without severe systolic dysfunction. *JAMA Cardiol* 2018;**3**. doi:10.1001/jamacardio.2018.1049

64 Chatterjee NA, Tikkanen JT, Panicker GK, *et al.* Simple electrocardiographic measures improve sudden arrhythmic death prediction in coronary disease. doi:10.1093/eurheartj/ehaa294

65 Lee DC, Albert CM, Narula D, *et al.* Estimating Myocardial Infarction Size With a Simple Electrocardiographic Marker Score. *J Am Heart Assoc* 2020;**9**. doi:10.1161/JAHA.119.014205

66 Panicker GK, Narula DD, Albert CM, *et al.* Validation of electrocardiographic criteria for identifying left ventricular dysfunction in patients with previous myocardial infarction. *Ann Noninvasive Electrocardiol* 2021;**26**. doi:10.1111/anec.12812

67 Cheng A, Dalal D, Butcher B, *et al.* Prospective observational study of implantable cardioverter-defibrillators in primary prevention of sudden cardiac death: study design and cohort description. *J Am Heart Assoc* 2013;**2**. doi:10.1161/JAHA.112.000083

68 Wu KC, Wongvibulsin S, Tao S, *et al.* Baseline and Dynamic Risk Predictors of Appropriate Implantable Cardioverter Defibrillator Therapy. *J Am Heart Assoc* 2020;**9**:e017002. doi:10.1161/JAHA.120.017002

69 Faxén J, Jernberg T, Hollenberg J, *et al.* Incidence and Predictors of Out-of-Hospital Cardiac Arrest Within 90 Days After Myocardial Infarction. *J Am Coll Cardiol* 2020;**76**:2926–36. doi:10.1016/j.jacc.2020.10.033

70 Bergau L, Willems R, Sprenkeler DJ, *et al.* Differential multivariable risk prediction of appropriate shock versus competing mortality - A prospective cohort study to estimate benefits from ICD therapy. *Int J Cardiol* 2018;**272**:102–7. doi:10.1016/j.ijcard.2018.06.103

71 Seegers J, Vos MA, Flevari P, *et al.* Rationale, objectives, and design of the EUTrigTreat clinical study: A prospective observational study for arrhythmia risk stratification and assessment of interrelationships among repolarization markers and genotype. *Europace* 2012;**14**:416–22. doi:10.1093/europace/eur352

Tables

Table 1: Overview of the datasets included in our study. The population column shows whether the data are of type "acute MI cohort", "prior MI cohort", "ICD cohort", or "randomised controlled trials", as defined in the methods section.

Dataset ([reference])	Dataset Includes MRI?	N	SCD and its proxies	Endpoints				Basic clinical characteristics		
				SCD or its proxy N (%)	Death other N (%)	Median follow-up (months)	Incidence rate of SCD or its proxy per 100 person-years	Age Mean (SD)	Sex Males (%)	LVEF (%) Mean (SD)
Aston University Research Database ([24])	Yes	805	SCD/ VT/ VF/ FAT	91 (11.3%)	234 (29.1%)	54.1	2.5	66.5 (12.2)	634 (78.8%)	42.1 (16.8)
Finnish Research Database (ARTEMIS) ([55];[56])	No	982	SCD/SCA	41 (4.2%)	152 (15.5%)	104.0	0.5	66.5 (9.2)	698 (71.1%)	61.8 (12.1)
EU-CERT-ICD Retrospective part ([57])	No	2,006	FAS	259 (12.9%)	271 (13.5%)	32.0	4.5	64.2 (10.1)	1,738 (86.6%)	26.4 (5.8)
Centro Cardiologico Monzino Registry	Yes	845	SCD/ SCA/ FAT/ VT	87 (10.3%)	76 (9.0%)	32.1	3.5	65.4 (11.0)	725 (85.8%)	34.0 (10.1)
DAI-PP pilot registry ([58])	No	1,580	FAT	367 (23.2%)	143 (9.1%)	29.6	8.4	61.7 (10.6)	1,405 (88.9%)	27.9 (5.5)
Helios Hospital EHR	Yes	452	FAT	134 (29.6%)	13 (2.9%)	22.8	10.2	65.0 (9.8)	400 (88.5%)	28.2 (5.7)
ISAR-RISK ([8])	No	3,821	SCD	82 (2.1%)	411 (10.8%)	56.7	0.6	62.9 (12.6)	2,833 (74.1%)	52.2 (13.1)
Israeli National ICD Registry {[59]}	No	753	FAT	54 (7.2%)	66 (8.8%)	28.4	2.8	64.8 (10.2)	685 (91.0%)	28.3 (6.6)
MADIT II Randomised Trial ([9]; [60])	No	1,231	SCD/ VT/ VF/ FAT	218 (17.7%)	122 (9.9%)	16.9	11.3	64.5 (10.4)	1,040 (84.5%)	23.2 (5.4)
MADIT RIT Randomised Trial	No	708	FAT	97 (13.7%)	26 (3.7%)	16.2	10.2	60.6 (12.5)	538 (76.0%)	26.5 (6.6)

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Dataset ([reference])	Dataset Includes MRI?	N	SCD and its proxies	Endpoints				Basic clinical characteristics		
				SCD or its proxy N (%)	Death other N (%)	Median follow-up (months)	Incidence rate of SCD or its proxy per 100 person-years	Age Mean (SD)	Sex Males (%)	LVEF (%) Mean (SD)
Nancy Research Database ([61])	Yes	100	FAT	36 (36.0%)	21 (21.0%)	57.9	7.6	58.2 (10.1)	86 (86.0%)	27.0 (5.6)
Olomouc Research Database	No	818	FAT	178 (21.8%)	116 (14.2%)	21.9	9.9	67.8 (10.1)	636 (77.8%)	30.7 (6.9)
PRE-DETERMINE ([62];[63–66]	Yes	5,781	SCD/ SCA/ FAT	256 (4.4%)	1,069 (18.5%)	89.0	0.7	64.2 (11.0)	4,401 (76.1%)	51.4 (10.6)
PROSe-ICD ([67]; [68])	No	394	FAT	62 (15.7%)	118 (29.9%)	48.3	4	64.3 (10.3)	335 (85.0%)	24.1 (6.8)
PROSe LV Structural Predictors Imaging Sub-Study ([67]; [68])	Yes	155	FAT	44 (28.4%)	47 (30.3%)	70.8	5.1	60.6 (11.0)	130 (83.9%)	25.7 (7.0)
SCD-HeFT trial ([11])	No	1,115	SCD/ SCA/ VT/ VF/ FAS	233 (20.9%)	215 (19.3%)	33.1	7.7	61.8 (10.6)	945 (84.8%)	24.6 (6.6)
Silesian Research Database	No	648	SCD/ FAT	25 (3.9%)	108 (16.7%)	55.0	0.9	64.2 (10.5)	432 (67%)	46.1 (8.6)
Swedish Heart Registry ([69])	No	175,573	SCA/VT/VF	3,239 (1.8%)	51,523 (29.3%)	48.1	0.4	71.0 (12.3)	114,352 (65.1%)	50.7 (11.7)
EU-Trig-Treat ([70]; [59])	No	115	FAS	16 (13.9%)	17 (14.8%)	44.6	3.9	65.8 (10.5)	99 (86.1%)	32.5 (9.6)
***Total**	-	**197,882**	**SCD/ SCA/ VT/ VF/ FAS/ FAT**	**5,519 (2.8%)**	**54,748 (27.7%)**	**48.1**	**0.6**	**70.2 (12.4)**	**132,112 (66.8%)**	**49.2 (12.9)**

* Excludes entries for the DO-IT registry (N=570) and Western Denmark Heart Registry (N=~28,000), which are analysed remotely.

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figures

Figure 1: Schematic of the inclusion and exclusion criteria, which are applied to each dataset within the PROFID project.

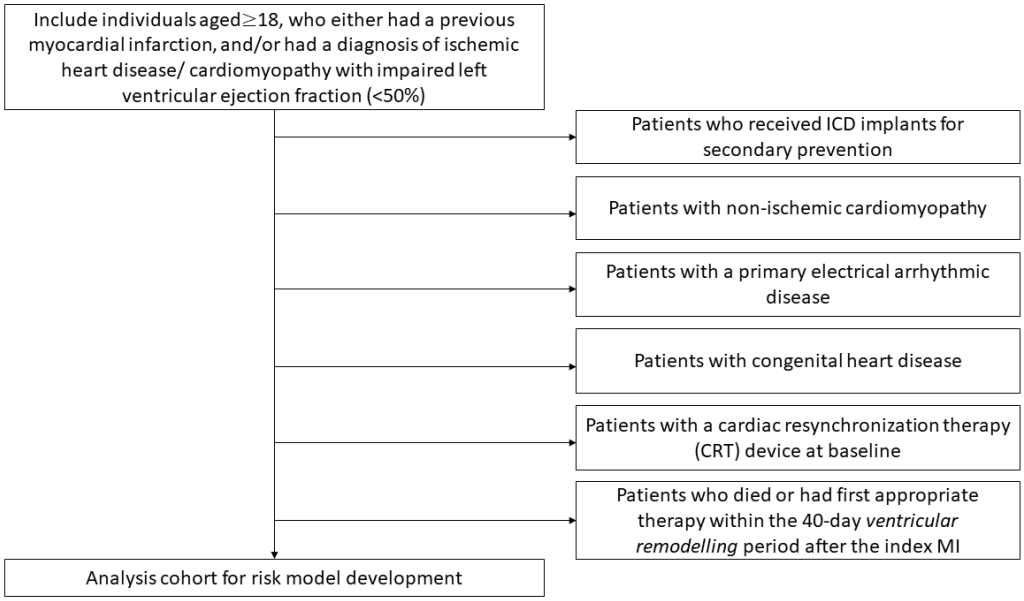
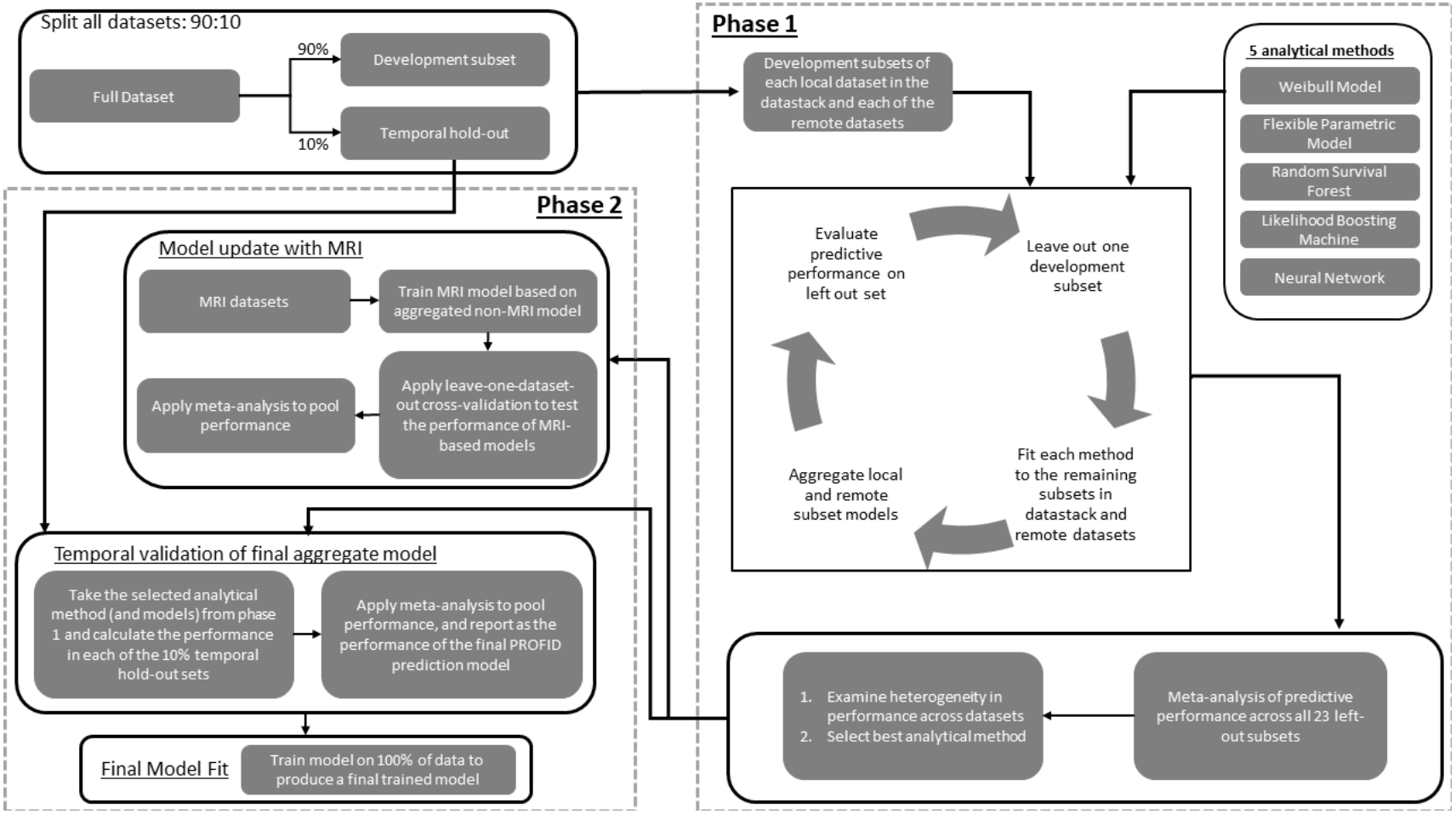


Figure 2: Graphical representation of our modelling approach.



medRxiv preprint doi: https://doi.org/10.1101/2021.07.12.21260002; this version posted July 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .