# Outcome Prediction

Buntheng Ly, Mihaela Pop, Hubert Cochet, Nicolas Duchateau, Declan O'regan, Maxime Sermesant

▶ **To cite this version:**

Buntheng Ly, Mihaela Pop, Hubert Cochet, Nicolas Duchateau, Declan O'regan, et al.. Outcome Prediction. AI and Big Data in Cardiology : A Practical Guide, Springer International Publishing, pp.105-133, 2023, 978-3-031-05070-1. 10.1007/978-3-031-05071-8_6 . hal-04212068

**HAL Id: hal-04212068**
**https://hal.science/hal-04212068v1**

Submitted on 25 Mar 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This is a pre-print version. The final document will be available at springerlink.com
DOI = https://doi.org/10.1007/978-3-031-05071-8_6

# 6 Outcome Prediction

***Mr Buntheng Ly***[a]

***Dr Mihaela Pop***[a,b]

***Prof Hubert Cochet***[c,d]

***Dr Nicolas Duchateau***[e,f]

***Prof Declan O'Regan***[g]

***Dr Maxime Sermesant***[a,*]

[a] *INRIA, Université Côte d'Azur, Epione team, Sophia-Antipolis, France.*

[b] *Sunnybrook Research Institute, Toronto, Canada.*

[c] *Department of Cardiovascular Imaging, Hôpital Cardiologique du Haut-Lévêque, CHU de Bordeaux, Pessac, France.*

[d] *IHU LIRYC, Université de Bordeaux-Inserm U1045, Pessac, France.*

[e] *Univ Lyon, Université Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France.*

[f] *Institut Universitaire de France (IUF), France.*

[g] *MRC London Institute of Medical Sciences, Imperial College London, United Kingdom.*

[*] *Corresponding author.*

**Authors' contribution:**

- Introduction, Opinion: DO.
- Main chapter: BL, MP, HC, MS.
- Tutorial: ND.

## Abstract

This chapter focuses on how we can best predict the future health of patients, known as prognosis. This encompasses areas such as risk prediction and predicting response to treatment. A clinical opinion piece summarises the role of prognosis in clinical care and highlights the areas where AI has already had an impact in this area. The technical section summarizes the state-of-the-art in outcome prediction, focusing on three clinical applications as exemplars: predicting response to cardiac resynchronization therapy (CRT), predicting outcome in atrial fibrillation and risk stratification in ventricular arrhythmia. A practical tutorial reinforces these concepts by taking the reader through a simple outcome prediction task based on cardiac morphology. The closing clinical opinion piece highlights areas where AI could impact prognostic tasks in the future.

**Keywords:**

prognosis, risk prediction, outcome prediction, cardiac resynchronization therapy, CRT, atrial fibrillation, ventricular arrhythmia, Kaplan-Meier curve

**Learning Objectives:**

At the end of this chapter you should be able to:

- *O6.A Compare and contrast traditional and AI based methods for outcome prediction in cardiology*
- *O6.B Explain some ways in which AI models can be used to predict response to cardiac resynchronization therapy, either using supervised or unsupervised formulations*
- *O6.C Describe how AI can be used to predict outcomes of atrial fibrillation*
- *O6.D Explain how AI can assist in risk stratification in ventricular arrhythmia*

## Clinical Introduction

Outcome prediction is a critical part of clinical decision making in cardiovascular disease. Accurate assessment of a patient's risk and timing of future events informs the choice of evidence-based prevention and treatment [298]. Imaging plays a pivotal role in risk stratification by visualising disease status, assessing disease trajectory and evaluating response to therapy. An example is the use of coronary artery calcium scoring (see also Chapter 4, page 100). as a semi-quantitative test for measuring calcified coronary artery plaque that can be of value in risk stratifying patients for future cardiovascular disease endpoints including guiding decisions about statin therapy in selected groups [146]. Although calcium scoring is a simple and highly reproducible test it doesn't account for prognostically important variations in regional distribution, intensity characteristics, or lesion-specific features [55]. A similar pattern of limitations emerges when using imaging to identify predisposing substrates and triggers associated with sudden cardiac death (SCD). Implantable cardioverter-defibrillators (ICD) are the most effective approach to primary prevention of SCD, and current guidelines regarding device implantation are based on an imaging-derived LV EF $\leq$35% [315]. However, the majority of out-of-hospital cardiac arrests occur in patients with only mild to moderate dysfunction who might be denied an ICD on current best practice [371], and so the reliance on single parameter thresholds fails to identify many of those who would benefit from the intervention [395].

Risk prediction guidelines draw insight from large-scale clinical studies through linear regression modelling of conventional explanatory variables, but this approach does not embrace the dynamic physiological complexity of heart disease [181]. Even objective quantification of heart function by conventional analysis of cardiac imaging relies on crude measures of global contraction that are only moderately reproducible and insensitive to the underlying disturbances of cardiovascular physiology [87]. In routine practice observer-driven pattern recognition is also used to guide classification introducing value from expertise but at the expense of objectivity and standardization [104]. Discretising severity into subjective categories may facilitate interpretability but incurs a loss of predictive power especially when building risk models [26]. Even consensus guidelines on patient management, for instance the investigation of stable chest pain, may substantially diverge when different assumptions, biases and inferential models drive their design [12].

The growing abundance of digital medical imaging linked to electronic health records presents an opportunity to develop prediction models that fully exploit biologically rich and diverse datasets at scale. Systematic quantification and evaluation of novel prognostic features could be transformative in the ambition for delivering "personalized medicine" tailored to individual characteristics including both phenotypic and genotypic profiles [379]. However, despite the exponential growth of machine learning approaches for prediction and classification tasks in healthcare [262], the safe and timely translation into clinically validated and regulated systems has proved to be challenging [192]. Systematic reviews of machine learning-based cardiovascular risk prediction have revealed inconsistent reporting, study heterogeneity and poor methodology [203]. In medical imaging applications of machine learning there are relatively few prospective or randomized trials, and independent external validation is scarce, increasing the risk of reporting biased performance estimates [278, 240].

Coordinated national and international efforts to enhance health interactions through access to large scale data and advanced analytics are accelerating the pace of prognostic algorithm development. Examples include community studies such as the 500,000 participants of the UK Biobank of whom 20% are being recalled for CMR [237], and the German National Cohort of 200,000 individuals including 30,000 with imaging [16]. Guidance is also emerging around the use of open data standards for healthcare informatics platforms to enable computable biomedical data to be discovered, analysed and evaluated in a trusted environment [346]. Here there is a growing role for federated learning architectures, where data are not exchanged, to address privacy concerns and provide access to heterogenous samples [329]. The most pressing bottleneck to progress is developing high-quality harmonized medical image data resources that have a robust ground truth coupled with active linkages to health events [157]. While the focus of the first wave of radiology AI applications has been on lesion detection, it is machine learning to guide risk stratification, assess treatment responses and perform outcome prediction that will be at the forefront of delivering actionable insights into clinical care [291].

Meaningful risk stratification must inform evidence-based management. For instance, an attractive target for better outcome prediction is where prognostically-rich data are not fully exploited by conventional analyses and any re-classification of risk group leads to a change in management [47]. Such individual-level

modelling requires clinical studies that capture how disease and treatment responses vary over time [351]. An advantage over developing sophisticated new image biomarkers of disease is that outcome prediction is readily interpretable, but it remains crucial to inform clinicians what features were important in the classification and being able to frame the results with a level of confidence. Where machine learning is brought closer to clinical decision making it is also vital to fully understand the role of human factors in such an unfamiliar cognitive environment – both for medics and patients. While the majority of patients currently support doctors using AI in the cardiovascular healthcare sector that confidence is easily lost and far more needs to be done to include stakeholders in setting priorities, ensuring trustworthiness, and addressing health inequalities [60].

## Overview

Following Chapter 5 on diagnosis, this section develops another classical problem in medical data analysis where AI has a strong role to play: outcome prediction. To help the reader appreciate the impact of AI in this field, the approaches taken by more traditional outcome prediction methods are first summarized. It is shown how methods for predicting outcome can be framed in different ways, and can make use of a wide range of disparate data sources. In particular, outcome prediction is often presented as a problem that can be addressed using a supervised learning formulation, given that in most cases labels can be taken into account (for example, the time to a negative event such as death or re-hospitalization, or even encompassing primary and/or secondary endpoints). However, this section also presents how an unsupervised formulation could help in some exemplar applications. This point of view is illustrated further in the hands-on tutorial accompanying this chapter (see page 168).

## Current Clinical Methods to Predict Outcome

Outcome prediction models of a disease or its recurrence following treatment are extensively used in clinical practice, medical research and public health [77]. In this regard, the ability to predict continuous or binary outcomes in patients with cardiovascular disease (CVD) has the potential for accurate identification of risk factors, stratification, superior treatment planning, as well as informed decision making [98, 323]. Specifically, modelling the outcome of arrhythmia-related cardiac diseases (such as atrial fibrillation, ventricular arrhythmia and heart failure) requires not only the selection of precise variables to accurately identify the critical predictors, but also to execute meticulous adjustments for time dependencies among treatments and responses [103]. Prior to the recent introduction of AI-based prediction methods, these prediction outcomes were modelled using classical statistical approaches, which are briefly outlined below along with associated terminology.

For cardiac arrhythmia-related conditions, survival data (i.e. the period from a specific time point to an event of interest [56]) refers to the time from arrhythmia episode or heart failure diagnosis to death or to any time-dependent phenomenon such as arrhythmia-free survival (i.e. the time until arrhythmia relapses). To understand arrhythmia-related survival data, one can generate

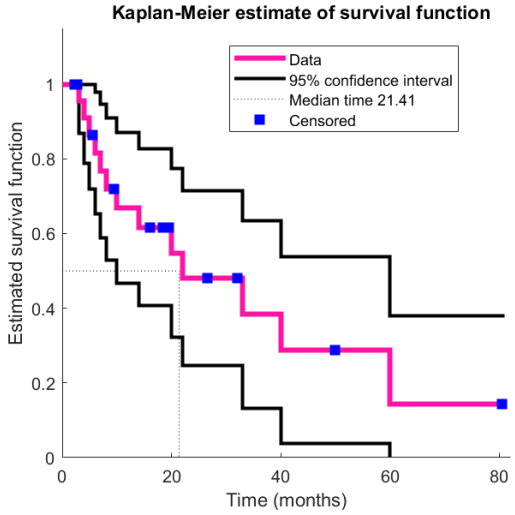



Figure 6.1: Kaplan-Meier curve to estimate survival probability (using virtual data and freely available code in [64])

a Kaplan-Meier (K-M) survival curve by representing time (days, months or years) on the horizontal axis and the calculated survival probability on the vertical axis. An example of a K-M curve is illustrated in Figure 6.1 using virtual data and freely available source code [64]. Explicitly, for each time corresponding to an event, a new value for the K-M curve is calculated by dividing the number of events that have occurred by the number of patients remaining at risk at that time, and then this new value is used to calculate the survival probability [141] and confidence interval. The censored data refers to incomplete data, such as in the case of a patient dropout from the study during the follow-up time. The so-called 'risk' is defined as being the probability of an event happening over a period of time. Should this risk vary over time, one can estimate the risk at a particular time point by calculating a new parameter named 'hazard'.

Typically, regression models or Cox proportional hazards models are employed for comprehensive analysis of the survival data. The Cox hazard model [175] relates the log hazard ratio to a linear predictor of one or multiple explanatory variables and is considered 'semi-parametric', meaning that there is no requirement to parameterise the underlying survival distribution. Cox regression models have been widely applied to predict the outcome of

abnormal heart rhythm conditions, although most of them cannot give information about when dangerous arrhythmic events might occur (or reoccur after therapy) within the following 1 year to 10 years. Novel risk prediction models can express results in a more specific time scale [370]. Although limited, data-based multi-variable statistical models correlate better with the actual patient outcomes compared to the predictions given by clinical experts, especially given the inter-physician variability.

The development of robust tools for primary and secondary outcome predictions is of great importance for all cardiovascular applications. Let us consider for instance the case of atrial fibrillation (AF), which is the most prevalent arrhythmia condition and is associated with life-threatening complications (e.g. embolic stroke, co-existence with heart failure, dementia) and death [427]. These complications and potentially fatal events lead to a considerable morbidity and mortality, posing a financial burden on the healthcare system. Notably, more than 30 million people worldwide suffer from AF, hence the considerable clinical interest to predict: the outcomes prior to the intervention; incident or recurrent AF after ablation; and the progression from sudden/paroxysmal to persistent or permanent AF. Despite a relatively high acute success rate of radiofrequency (RF) catheter ablation therapy, the outcome prediction of long-term AF recurrence during follow-up remains challenging. Using regression with multiple variables, various clinical scores can be calculated, such as the APPLE score (using: age, persistent AF, imPaired eGFR, left atrium LA, ejection fraction) at baseline with rhythm outcomes documented using 1-week monitoring with Holter device [202], or the MB-LATER score (using: male gender, bundle branch block, LA, AF type, early recurrences) 3 months after ablation, although the predictive ability of these scores may appear modest [311].

Other clinical prediction methods of AF outcome rely on tedious classification of signals recorded by the common 12-lead electrocardiogram (ECG) [212], the amount of atrial fibrosis identified by CMR imaging [75], or CT imaging-defined atrial shape statistics [174]. However, the former predictor requires a substantial amount of dedicated time and resources in order to process a large number of ECG signals, whereas the latter predictor is limited by the relatively poor spatial resolution of the data acquired in clinics and by the fact that most image-based segmentation methods still lack thorough validation. Thus, a consequence of using more sophisticated prognostic and risk prediction methods for primary or secondary outcome predictions

is that the number of input variables becomes significant. This leads to complex regression models, potential bias, and difficulties in assessment of model performance via calibration and discrimination measures. However, the utility of using calibration (i.e. overall performance and goodness of fit) and discrimination (i.e. predictive values, ROC curve) measures is uncertain, and cannot guarantee the robustness of the prediction model and its overall contributions to the net benefit and cost effectiveness of the study.

Equally important, it should be underlined that current approaches to predict outcomes strongly depend on: statistical assumptions of the model employed; data source and standardization; sample size in large cohorts of patients; misinterpretation of scores; cumbersome long-term survival analysis (including missing data at follow-up and/or unexpected mortality); as well as on multivariables in the model (clinical and therapy-related taken at baseline), which altogether complicate the analysis [77].

## AI-based Methods to Predict Outcome

To address the limitations of traditional methods employed for clinical outcome predictions in CVD patients, recently developed tools using machine learning concepts either based on agnostic approaches or on data-driven models empirically optimized have been proposed. These can partially overcome the issues associated with the traditional regression-based prediction methods. However, some initial machine learning-based methods (e.g. SVM or random forests, see Chapter 5, page 121) did not prove to be sufficiently superior, especially when looking at the ROC curve (more specifically, AUC. see Chapter 2, page 35) as a criterion for comparisons between the outcome predicted by these machine learning methods vs. traditional regression methods [85]. Thus, better approaches are still needed and these should be able to handle multi-variables input as well as complex relationships between inputs and output prediction, while being customized for the data specific to a particular clinical study. In this context, several novel AI-based methods have been developed to accurately and robustly predict complex clinical outcomes, as illustrated in this chapter for arrhythmia and dyssynchrony.

Outcome prediction models must be able to forecast a future event based on the pre-recorded patient's descriptors. As shown in Figure 6.2, these descriptors can include: specific image-based biomarkers (e.g. amount of fibrotic scar or wall thickness, atrial/ventricular shape, indices like ejection

fraction and strain); physiological ECG signals or blood pressure; as well as clinical baseline descriptors such as gender, race, phenotype, etc. Based on these features, the AI-based models are optimised to group the patients into specific outcome classes. Technically the model can be defined in a similar way to diagnosis models (see Chapter 5), but the main difference is the delay between the descriptor registration and the desired endpoints, where the complexity of the ground truth outcome and the evolution of the clinical descriptors can be recorded and exploited through time.

The clinical outcome can be complex (depending on the disease and/or therapy of interest) and can include: the acute success rate or response to a specific therapy; the mid-to-long term survival rate following the therapy; other events such as intervention-related complications, worsening of already existing comorbidities, or sudden cardiac death (SCD).

The class output of the prediction model can be defined as binary, based on the patient's status at a specific time point, or can be diversified into more classes to include the status or evolution at each follow up point, for instance the event/death occurrence in the 1st, 2nd or 3rd year, etc. Furthermore, while accurate AI-based outcome predictions based on pre-therapy/follow up descriptors would be beneficial for clinical decision making and therapy planning, knowledge of the dynamic evolution of these descriptors post-therapy could provide valuable insights for modelling an optimized response.

Integrating the descriptors at each follow up into the AI outcome model has potential not only for accurately predicting the patient status at the next follow up point, but also for a better understanding of the relationship between descriptor volatility and the eventual clinical outcome.

In the following subsections, we provide three applications of AI-based methods implemented for modelling outcome predictions for distinct pathological cases, namely: heart failure; atrial fibrillation; and ventricular arrhythmia. The scope of this chapter is not intended to be an exhaustive review of all the AI-based methods for outcome predictions; thus, the methods presented below are meant to provide representative examples from our field of expertise and to illustrate how supervised and unsupervised AI approaches (introduced in Chapters 2 and 5) could be integrated into clinical outcome prediction pipelines.

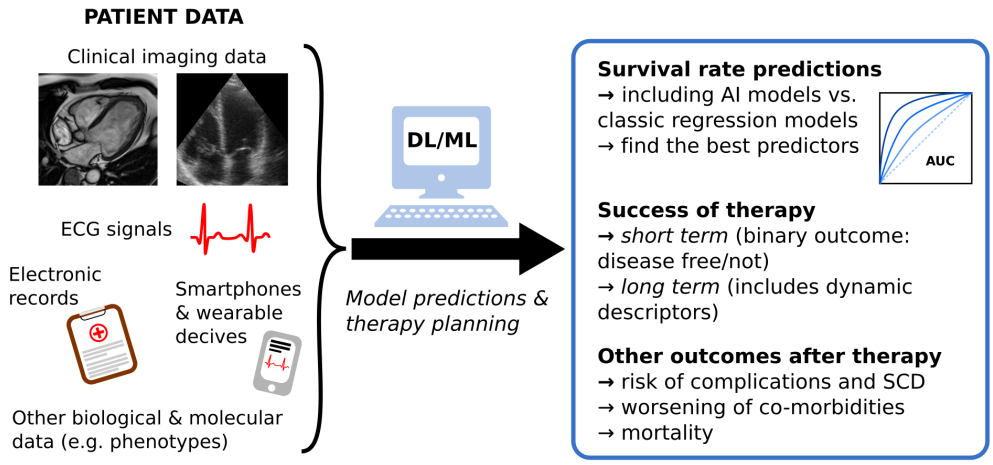



Figure 6.2: Example of generic pipeline for building AI-based models to predict survival rate or therapy outcome for cardiac applications.

## Application: Prediction of Response Following Cardiac Resynchronization Therapy (CRT)

Cardiac Resynchronization Therapy (CRT) involves the implantation of a biventricular pacing device in selected patients with mild to severe systolic heart failure (HF) to address the symptoms of HF and to reduce HF hospitalizations. The pacing restores a synchronous beating of the right and left ventricles, improving the overall biomechanical function of the heart and, consequently, the ejection fraction (EF). According to American Heart Association (AHA) and European Society of Cardiology (ESC) guidelines, two official evidence-based guidelines for HF management, CRT provides a clear-cut benefit to patients with reduced LV ejection fraction ($< 35\%$), prolonged QRS duration ($> 150$ms), left bundle branch block (LBBB) morphology, and in sinus rhythm, who are still at risk of advanced HF progression despite receiving optimal medical treatment. Response to CRT corresponds to the degree of LV remodelling documented in the imaging, usually by quantifying the reduction in the LV end systolic volume. However, the evidence for positive CRT response becomes less clear when the QRS duration is between $130 - 150$ms, non-LBBB morphology or with AF patients, since the recommendations start to deviate between the two guidelines [396]. In addition, depending on the current selection criteria, between 20% to 30% of patients who underwent CRT were reported as non-responders [314]. While strate-

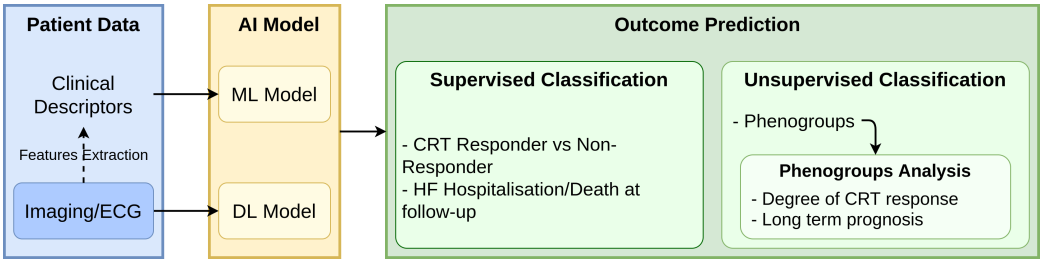



Figure 6.3: Formulation of AI-based model prediction for CRT outcome.

gies to improve CRT response might also involve the improvement of CRT technology and post-implant care, pacing optimization and patient selection both still play a major role in limiting unnecessary implants and correctly assigning patients to appropriate treatment.

Moreover, while the current recommendations are based largely on LV ejection fraction, QRS duration and morphology, several clinical trials have demonstrated that patient response to CRT also depends on demographic and clinical characteristics as well as on the electrical and mechanical function of the heart [314]. Thus, the interest in CRT patient assessment has started to shift towards the inclusion of imaging data. This is where AI methods have made their way into CRT response prediction, thanks to their ability to integrate and interpret diverse and heterogeneous data involved in treatment personalization for superior CRT outcome. CRT outcome prediction using AI-based models can be formulated as a supervised or unsupervised problem, as shown in Figure 6.3. It should be noted that recent developments suggest that AI-based models outperform conventional clinical methods [42]. The following subsections will provide selected application examples of AI-based models built for CRT outcome prediction. As AI is an emerging technique, the reader is advised to seek contemporary reviews such as, for instance [42], for a more comprehensive review of AI methods used in CRT.

### Supervised Prediction of CRT Response

Supervised outcome models are trained to predict the endpoint according to the input descriptors, which, in the context of CRT outcome, can provide a direct answer as to whether the patient would benefit from the therapy. To train such models, datasets comprised of the ground truth endpoint are required. The primary endpoint of CRT clinical trials usually entails ei-

ther death from any cause or nonfatal HF events. However, such datasets are usually not publicly available and their acquisition implies large clinical trials that span over multiple years. In addition, the retrospective nature of these datasets means that the available clinical descriptors or imaging data are limited by the study protocol. This can limit researchers' ability to investigate novel biomarkers since the required data may not have been recorded. Therefore, to facilitate these studies, classification tasks are usually simplified and focused on patient response to CRT. The labels "responder" or "non-responder" are assigned to the patients who showed significant LV remodelling as quantified by the reduction in the end systolic volume (between >= 10% and >= 15%) at 6-month post-operation follow up. This predictor was shown to be a strong indicator of lower long-term mortality and HF events [431].

Both machine learning and deep learning methods have been shown to provide additional predictive value over the metrics used in current clinical guidelines (LV ejection fraction, QRS duration and LBBB) [304, 317, 384, 185]. In addition to these metrics, machine learning methods are able to exploit detailed cardiac motion data for outcome prediction (i.e. response to CRT) [304, 185]. For example, a random forest-based machine learning model was able to achieve a higher AUC score (0.74 compared to the log regression model 0.67) [185]. Furthermore, owing to their ability to process large multidimensional input data (i.e. the imaging data), deep learning methods are capable of making accurate predictions from the LV and RV segmentation masks of CMR images [317]. Using such masks of heart motion through the cardiac cycle phases, a deep learning model can learn to predict patient CRT response without the need for feature extraction or additional clinical descriptors.

The binary evaluation, "responder" vs. "non-responder", using a single cut-off value of the LV end systolic volume, might not accommodate all the possible outcomes and the subtlety of every patient's response to CRT. Moreover, the categorization is even more heavily impacted by the poor reproducibility of the serial LV end systolic volume measurements. To mitigate this issue, a "super-responder" class can be considered in the classification model, which provides information on the patients most likely to gain strong benefit from the therapy. Given appropriate data for supervised training, machine learning can be used to predict super-response as well as just response to CRT [303]. In addition, in studies based on long-term CRT clinical trials , ma-

chine learning methods can be used to provide more prediction details than simply response or even super-response. Patient survival through the follow-up period could also be framed as the model output [384]. In this case, the output prediction may be split into different classes according to the patient survival post-CRT therapy, which could provide better insight into the clinical evolution of the pathology, offering potential benefits to decision making and planning strategy.

An AI classification model usually predicts the output as a value between 0-1 for each class, which is usually regarded as the probability that the inputs belong to the specific class. However, most incorrect predictions are still associated with a high probability. Enforcing the model training to be uncertainty-aware [101] could provide additional information when analysing the model output. The integration of uncertainty into model predictions also allows including the variability of clinical data in the prediction (image and segmentation quality, incomplete clinical variables, etc.). This variability can be more prevalent in routine clinical care compared to data from clinical trials. Complementing the model output with uncertainty information could be extremely useful to ensure clinical adoption of the AI model as a decision support tool.

#### Unsupervised Prediction of CRT Response

While supervised models may be bounded by the output class, unsupervised models do not require any output label for the learning phase. From the perspective of better characterizing the patient outcome, the main objective behind this family of methods is to fit the patients (as represented by the input characteristics) into different archetypal subgroups or phenogroups, according to the similarity of their characteristics and not based on already existing labels. These phenogroups can be then interpreted by analysing their differences in outcome and patient characteristics.

Being agnostic to the potential labels makes an unsupervised approach capable of identifying two phenogroups of patients based on differences in clinical characteristics and long-term prognosis [132]. The survival analysis of the unsupervised phenogroups highlights the distinction in survival rate between the two populations. The phenogroup analysis can also be related back to the input characteristics. For example, in [132] one phenogroup was found to have higher numbers of CRT responders as well as certain input fea-

tures (apical rocking, septal flash) whereas a second phenogroup featured signs of advanced HF (RV dysfunction, kidney failure and biventricular dilation).

The number of phenogroups in the unsupervised model is not limited to two, and several profiles can be generated to account for the spectrum of possible outcomes following therapy. The optimum number of phenogroups can be calculated independently to the endpoint by maximising the distance of the phenogroups [132]. This optimum number can also be set to maximise the statistical significance of the phenogroups to a desired endpoint. Note that while the number of the phenogroups may be biased toward a ground truth label, the optimization does not take the label into account and thus the population is still grouped according to the correlation of the input characteristics. An initial number is chosen for the first training, then the model is retrained with the new number of phenogroup(s), until the optimum condition is met. Up to 4 phenogroups can be defined based on the primary endpoint (death or non-fatal HF event), to account for the different levels of prognosis: the best, the worst and two in-between phenogroups [86]. The survival analysis of the population in each phenogroup proves the accuracy of the unsupervised model in grouping the patients likely to benefit or not from CRT.

Without prior knowledge, the unsupervised model was able to classify the CRT "responder" vs "non-responder" groups in a statistically significant manner with better accuracy than each single clinical descriptor alone [86]. It is also interesting to note that without supervision, the identified phenogroups actually correspond to specific mechanisms that can condition CRT (non-)response, previously described by clinicians based on their physiological knowledge [296]. Although the binary division of patients may appear too simplistic to accommodate for all the possible patient reactions to therapy and their long-term prognosis, the unsupervised model allows assigning a "risk profile" to the patient according to the common clinical descriptors, which could prove useful in CRT patient selection and clinical decision making. An additional benefit of unsupervised learning is the flexibility on the exploitable data. Although a significant number of known ground truth outcomes are required for accurate phenogroup analysis or at least better interpretation, unsupervised training can advantageously, by definition, be performed on databases with unknown or incomplete outcome labels. The lack of implicit classification loss during the optimization also limits the over-

fitting in small datasets compared to supervised methods [443].

Beyond the specific context of CRT, unsupervised phenogrouping approaches may be useful in the overall context of HF to unravel novel disease entities, knowing, for instance, that dilated cardiomyopathies are currently poorly classified. This would in turn allow more personalized patient management by selecting the drugs or interventions (e.g. CRT) most likely to be effective for the specific phenogroups.

## Application: AI Methods to Predict Atrial Fibrillation Outcome

Accurate prediction of primary and secondary outcome in case of arrhythmic events is a critical task for early prevention and selection of the most effective treatment. Atrial fibrillation (AF) is the most prevalent arrhythmia condition, which along with its associated comorbidities [209], represents a burden to healthcare systems and an increased risk of stroke and mortality to the patient, particularly for the ageing European and North American populations. Among the most important comorbidities are: coexisting HF, ischemic heart disease, hypertensive or valvular heart disease, and diabetes. With respect to AF management, the first line of therapy is anti-arrhythmic medication (e.g. beta blockers, calcium channel blockers) to control the heart rate, along with anti-coagulants that prevent blood clots and stroke. However, during prolonged treatment spanning over years, the anti-arrhythmic drugs are often associated with side effects (e.g. shortness of breath, dizziness, tiredness, slow heart rate, low blood pressure) and also affect over time the normal function of several organs (e.g. liver, kidney, thyroid, lungs). Other therapy options are cardioversion (to reset aberrant heart rhythms), and catheter ablation (to eliminate the atrial foci generating abnormal electrical impulses). Notably, among this spectrum of therapies, only catheter ablation is potentially curative. This minimally-invasive procedure is performed under imaging guidance and consists of the elimination of AF foci using thermal energy (e.g. radiofrequency and cryoablation), via an ablation catheter whose tip is precisely manoeuvred to destroy only tiny tissue areas harbouring AF sources. Several predictors of AF risk as well as of the outcomes prior to and following the therapy of choice have been clinically identified. Among the key predictors are: ECG signals (recorded in ambulatory, clinics or by wearable devices: Holter monitors and smart watches); biomarkers extracted

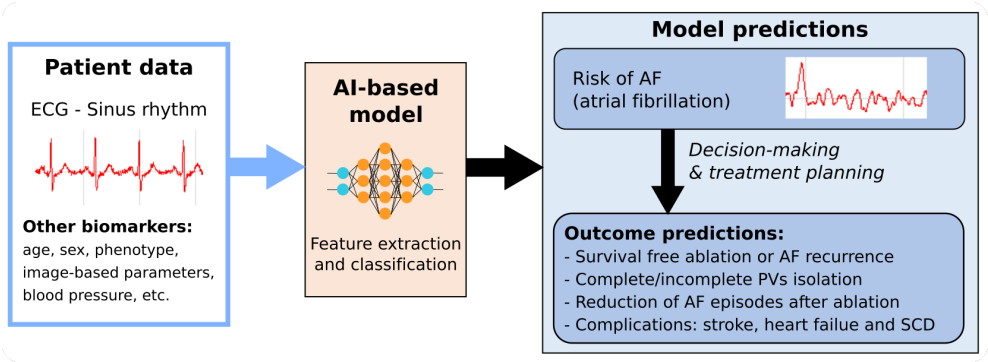



Figure 6.4: Example of AI-based pipeline to predict AF risk and outcome.

from clinical data (age, race, sex, phenotypes, image-based parameters such as the amount of fibrosis, atrial shape and size descriptors: surface area, anteroposterior diameter, biplane area-length volume) and information from patient electronic health records [400]. As described in more detail below, the features of these descriptors can be extracted and used by AI-based models to predict AF risk and important outcomes such as: whether the patients are free of AF, whether AF reoccurs following ablation but has reduced number of episodes; complete vs. incomplete pulmonary vein (PV) isolation during ablation; worsening of comorbidities, embolic complications (e.g. stroke), and overall mortality (see Figure 6.4).

The most important key descriptor of AF is the noninvasive ECG, a widely available monitoring measurement of cardiac electrical activity obtained by means of one or multi-surface electrodes. ECG is an established clinical diagnostic biomarker of abnormal heart rhythm (too fast, too slow or with irregular beats), which can be easily digitized and transferred for interpretation. The typical components of recorded ECG signals are: P wave (corresponding to atrial depolarization), QRS wave (ventricle depolarization), and T wave (ventricular repolarization). The QRS complex is often converted to a Fourier spectrum in order to observe dominating events and potentially lethal AF in the 0-20Hz frequency range [390]. Unfortunately, the signature detection of various types of AF morphologies needed for classifiers that feed traditional regression models requires experts for interpretation as well as dedicated resources to analyze large datasets, which are difficult to find.

These limitations have prompted research into versatile methods built on

adaptable deep neural networks that can deal with such large datasets, and on machine learning methods empowered to have learning abilities. While promising, the early machine learning-based prediction models used PCA, SVM or random forest methods, and employed intense preprocessing steps and noise removal before extracting relevant morphological features from the ECG signals (e.g. slopes, peaks, amplitude timings, etc.) [248]. However, modern convolutional neural network-based (CNN, see Chapter 3, page 74) models have the ability to use feature characteristics extracted directly from raw ECGs for automated analysis. In contrast to 2-D CNN models that are suitable to exploit image structure, deep learning-based models using 1-D CNNs are able to segment and classify the heart beats using ECGs. Each beat is labelled as normal or abnormal, enabling predictions of AF risk and stroke complications as outcome [421]. Other models use 12-lead ECGs recorded in sinus rhythm in order to find suitable patterns to predict incident AF [261], while recently developed CNN models can be trained with more than 1 million ECGs to accurately predict mortality as a primary AF outcome [323]. Complementary details on ECG analysis but not necessarily specific to AF can be found in Chapter 10, page 237.

Lastly, a notable recent breakthrough of AI-based methods for AF outcome prediction is in the area of remote monitoring technologies, where automatic AI algorithms have been applied to single-lead ECG traces obtained through mobile and smart watch-enabled recordings [344, 115]. It is envisioned that smart AI-based algorithms developed for consumer-or patient-facing applications (which are massively scalable) will soon completely exceed the capacity of human readers of ECGs. However, the utilization of these predictive models for AF risk and outcome is still hampered by the inconsistent quality of data collected in real-time fashion. This is mainly due to the sporadic poor quality of tracing and to noisy data, which might introduce bias and error in the correct interpretation and the model output.

To sum up, integrating complexity into AI-based prognostic models through multilayer deep learning models can result in rapid identification of ECG signal features and subtle patterns which are not typically recognizable by the human eye. Comprehensive and sizeable clinical datasets containing single-lead or multi-lead digital ECGs are also being linked to electronic health records (see Chapter 10, page 243), substantially contributing to the development and deployment of accurate AI models for AF risk and associated outcome prediction.

## Application: Risk Stratification in Ventricular Arrhythmia

Ventricular Arrhythmia (VA) is the most frequent event leading up to SCD, which is among the major causes of death in developed countries. The spectrum of therapies includes the delivery of electrical shocks to the heart via implantable cardioverter defibrillators (ICD) to prevent SCD, and catheter radiofrequency ablation as the potential curative treatment. Both therapies involve invasive and risky interventions; thus, the correct identification of patients at risk as well as the ablation targets (i.e. the discrete myocardial sites promoting arrhythmia) are crucial to prevent SCD and reduce surgery complications.

The ICD is an implantable device used to deliver appropriate electrical therapies (antitachycardia pacing or shock) to terminate the VA episode. The ICD implantation is applied preemptively to subjects identified as being at risk of developing potentially lethal VA. The objective of the therapy is to terminate the arrhythmic episode at the occurrence, but not to prevent its recurrence. The current recommendation for ICD patient selection for primary prevention relies largely on the LV ejection fraction value, which is a key clinical index measuring the relative change of LV volume between end diastole and end systole [19]. Unfortunately, current clinical strategies based solely on the LV ejection fraction lead to numerous nonessential implants, due to the fact that up to 3/4 of the selected patients would not receive any appropriate therapy within 5 years after the implantation [345]. In addition, the current strategies miss more than 80% of SCD victims whose LV ejection fraction is not severely altered.

Radiofrequency ablation is an electrophysiology procedure that eliminates the VA source (known as the 'substrate') using an electrical current delivered by an intracardiac catheter whose tip is maneuvered onto the target. Ablation is proposed for patients in the advanced stage, who have experienced multiple VA episodes and who often received multiple ICD therapies that were poorly tolerated. The objective is to modify the myocardial substrate on which arrhythmias occur, in order to prevent their recurrence. The main limitation of this treatment lies in the correct and exhaustive identification of the target. The current diagnosis strategy in the electrophysiology lab involves dangerous VA induction using a programmed electrical stimulation [19], which is invasive and time consuming, and suffers from a limited ability

to successfully induce arrhythmia and inaccessibility to the arrhythmogenic substrate location.

Therefore, accurate VA risk stratification is crucial for adapting the appropriate therapy for SCD prevention. Furthermore, the classification of VA patients should also be extended to the detection of specific arrhythmogenic areas for successful curative ablation interventions.

### The Machine Learning Approach

The poor performance of LV ejection fraction in patient selection for VA could be explained by the limitation of a single descriptor to predict a complex phenomenon such as VA. This further highlights the limitations of classical statistical analysis, which usually focus on the impact of single descriptors. In contrast, machine learning provides models that can capture more complex statistical relationships and integrate more disparate and multidimensional data.

There is no limit, technically-speaking to the number of variables that can be used as input to machine learning models. These could consist of demographics, medical history, medication therapy, laboratory results, and features extracted from ECG, imaging and clinical notes [42]. Ideally, since the model would be able to learn by itself to distinguish which inputs are useful or not through optimization, it is advisable to include all the available descriptors as input to avoid feature selection bias.

Machine learning-based models could also be used to integrate the evolution of the input variables in a dynamic way [418]. For example, the RF_SLAM (Random Forest for Survival, Longitudinal and Multivariate) model allows the integration of baseline descriptors (pre-ICD implant) and dynamic descriptors (post-implant). The updated clinical descriptors, at each follow up, such as the serial LV ejection fractions or the number of HF hospitalizations are integrated to the prediction model to provide new information concerning patients' biological response to the therapy and their survival rate. Such a dynamic model provides a better understanding of the relation between the evolution of the variables, allowing the flexibility needed in personalixed medicine. Lastly, the patient HF status post-therapy plays an important role in predicting patients' survival, while the the serial LV ejection fractions may not contribute substantially [418].

However, more input variables lead to more complex models, which would take longer to train and to run, and be more difficult to interpret. Moreover, external issues such as missing data or clinical practicality could be legitimate reasons to limit the number of input descriptors. Current statistical methods to reduce the number of input variables mostly rely on univariate and multivariate analysis for feature significance, where the variables with significant p-value ($< 0.05$) are selected [418]. With machine learning models, feature importance analysis can also be used for feature selection. First, the primary model is trained with all the available input variables, and then the top predictors are extracted to be used as the input to the secondary and main prediction model [403].

The feature importance analysis of the machine learning model is an analysis of the degree of importance of each input variable to the model decision. Feature importance algorithms can be model dependent, by inspecting the weights or coefficients of the trained model, or model independent, for instance using the permutation importance algorithm[25], which looks at the score decrease when a feature is absent. Understanding the importance of each input variable allows transparency and interpretability, which are required to escape from "black-box machine learning models" (see also Chapter 8, page 205, and Chapter 9, page 227), and help increasing trust and viability of the models in clinical practice. Features such as the presence of left bundle branch block, serum magnesium, antiarrhythmic drugs, LV scar size, and LV gray zone have been reported to be among the most influential clinical descriptors of VA [403, 418].

The integration of multi-modality data allows the machine learning model to play a vital role in personalized medicine. Understanding the importance of imaging features in the prediction of outcome, such as LV scar or gray zone, is also a pivotal step in radiomics, an emerging field that explores a large variety of quantitative features derived from medical images.

### Feature Extraction Before Learning

In VA risk prediction, the cardiac descriptors to be used as input for machine learning can be extracted from different types of imaging modalities, including echocardiography, CMR or CT imaging. Depending on the available

[25] https://scikit-learn.org/stable/modules/permutation_importance.html#id2

imaging data, these descriptors could be static or dynamic. Static features are extracted from the image captured at a specific moment of the cardiac cycle, in general end diastole and/or end systole. These can include anatomical features of the heart such as myocardial scar, myocardial thickness, or LV volume. Dynamic features are the features that define the movement of the heart and can be extracted from image sequences throughout the cardiac cycle. Features such as myocardial displacement, strain, or strain rate along the main anatomical directions of the heart can be extracted.

Feature extraction usually starts with image segmentation and tracking, which is generally performed fully manually or semi-automatically. From there, the above-mentioned features can be extracted by image processing. Deep learning methods can be used for robust automatic segmentation and even tracking in many cardiac imaging modalities (see Chapter 4, page 92). This allows fully-automated feature extraction, which is crucial to exploit large databases where manual extraction on all cases may not be feasible. Nevertheless, automatic segmentation models still need to be improved for some tasks on specific imaging modalities, such as LV myocardial scar delineation from late gadolinium enhancement CMR [189, 444]. In this case, manual segmentation or at least manual corrections on top of an automatic delineation is highly recommended.

Extracting relevant features from images plays a crucial role in refining the raw image data, which may not be adapted for a given machine learning-based outcome prediction model (using classification or regression). This also serves to obtain features that are more human understandable, thus allowing better interpretability of the prediction, and even more when this is combined with some feature importance analysis as mentioned above. There are still some limitations to this approach. First, the extracted features are usually grouped into regional and global features (namely, by averaging local values across a given region or the whole myocardium). Regional features might allow a certain flexibility into regional heterogeneity compared to global features, but they do not allow the assessment of finer heterogeneities within the region. Second, using the extracted features as inputs to a machine learning-based outcome prediction model does not explicitly provide spatial information to the machine learning model. The model would have to learn this spatial or temporal relation between the variables during its optimization, which represents an unnecessary extra step. In contrast, the spatiotemporal information would not be lost for CNN models, for instance,

which allow the direct used of imaging or ECG sequence data as input. Finally, the limited types of features usually extracted from imaging data can also lead to selection bias, namely only the same known features are studied, while other features (still present in the images) are ignored.

### Going Further With Deep Learning Models

Deep learning models are capable of predicting outcome without having to extract specific features from the images, as exemplified in many classification problems in computer vision, which fostered the popularity of CNNs. In healthcare applications, direct diagnosis can be obtained for a large range of domains including (but not limited to) dermatology, cancer or lesion detection, and fracture detection. Nonetheless, direct VA classification using a deep learning model working on the raw image data as input has not yet been reported. A potential reason for this is that it requires considering complex cardiac data, in 3-D or even 3-D+time, meaning additional complexity of input data for a classification model.

In the context of VA risk stratification, myocardial scar is considered a substrate leading to the VA mechanism. Electrophysiology assessment has linked myocardial scar with myocardial fibrosis, a pathological remodelling of the cardiac muscle [285]. The gold standard technique to visualise scar is late gadolinium enhanced CMR. The features of LV myocardial scar extracted from LGE CMR imaging have been shown to be determinant predictors for VA, already by themselves [198] or combined with other descriptors in a machine learning prediction model [418]. Although scar segmentation requires manual segmentation by an expert (until automatic methods reach acceptable performance on data from clinical routine), these works highlight the potential of deep learning models to use scar segmentation data for VA risk stratification.

CT imaging has also proven to be relevant for myocardial scar imaging, in the form of visualizing wall thinning, which is known to have similar electrophysiological properties as the scar region observed in LGE CMR images [201]. Thus, for CT imaging, a major objective would be to first quantify LV wall thinning using segmentation techniques, which would alleviate the extra task of segmenting scar within the myocardium, which is still challenging in CMR images. Moreover, compared to CMR images, CT images have better contrast and resolution, and image acquisition is better standardized across

imaging centres and scanner manufacturers. These conditions make the use of deep learning models for automatic segmentation attractive.

Once the LV myocardium is segmented, the 3-D structure can be further simplified by calculating myocardial thickness locally and projecting these values onto a 2-D Bull's eye representation of the whole LV. The flattening helps to reduce the dimensionality of the data and therefore the computations involved during learning. It also limits the effect of (zero) values outside the myocardium if the 2-D or 3-D data are considered as 2-D or 3-D images.

The Bull's eye flattening of the LV is inspired by the American Heart Association 17-segment model, which helps physicians to better understand the distribution of input values across the 3-D LV myocardium. These steps can be fully automated, meaning that outcome prediction studies can be performed on large datasets. Datasets of CT images are also easier to construct, due to the wider availability and inclusiveness of this imaging modality (i.e. it can be performed on patients with metal implants) compared to CMR imaging, which enhances the feasibility of building large prospective databases in the near future.

Through their optimization, deep learning models can learn the relationship between the extent, position, and heterogeneity of the wall thinning region and the patient's risk of arrhythmia. These models would have to learn to filter between pathological and physiological wall thinning, as observed at the base and apex of the LV, and thickness heterogeneities, which could be caused by the papillarity muscles and trabeculations. This has been recently shown to outperform predictions based on the LV ejection fraction [247].

### Explainability With Deep Learning Models

With "standard" machine learning models, input features generally involve previous design by the user and their relative importance can be studied to interpret the prediction. In contrast, deep learning models stand more as "black boxes" that transform the input data into a prediction, which limits human understanding of the model decision (see also Chapter 8, page 205, and Chapter 9, page 227). While their performance might be higher, the lack of explainability can clearly limit the trust from both physicians and patients, therefore making the integration of deep learning models in clinical practice harder to justify. To achieve some transparency with deep

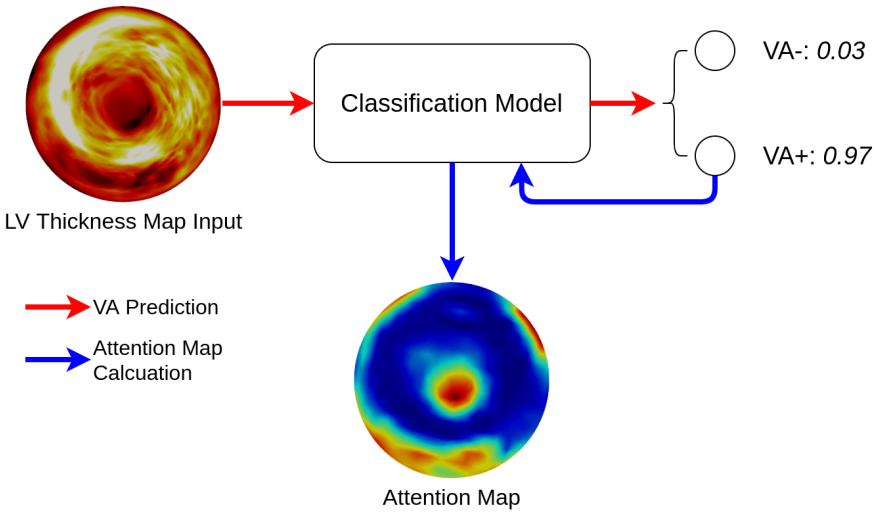



Figure 6.5: The Ventricular Arrhythmia (VA) prediction output and corresponding attention map [247]. The classification model classified the input thickness map as VA+ (with the score of 0.97). From the VA+ score, the GradCAM++ method [74] used gradient back propagation to generate the attention map, which highlighted the regions most influential to the model prediction.

learning classification methods, visual explanations of the prediction could be estimated, in the form of an attention map to answer the question *why did the model predict what it predicted?*.

To illustrate this, Figure 6.5 shows the attention map calculated with the GradCAM++ technique [74] (see Technical Note, below) from a positive VA prediction [247]. A trained classification model was used to classify the 2-D LV thickness map input, which provided two scores for VA+ and VA- (i.e. presence or absence of VA). Following the GradCAM++ method, the positive score was backpropagated to the"last", i.e. the deepest, convolutional layer to generate the classification attention map. We can observe that the map highlights the thinning region in the input, which allows clearer understanding to the model's prediction and further confirms the initial hypothesis that linked myocardial thinning with risk of VA.

**Technical Note**

The class activation mapping (CAM) method follows the fundamental assumption that $Y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k$, where $Y^c$ is the classification score of class $c$, $w_k^c$ is the weight for each specific feature map, $A^k$ is the feature map of $i \times j$ resolution from the last convolutional layer of $k$ filters and $i$ and $j$ stand for the row/column indices of each pixel. In other words, the classification score of class $c$ can be calculated as a linear multiplication of the global sum of the last convolutional feature maps $A_{ij}^k$ and the unknown weights $w_k^c$ for each feature map $k$. The class-specific attention map, for the spatial location $(i, j)$, can then be calculated using $L_{ij}^c = \sum_k w_k^c . A_{ij}^k$.

The weights $w_k^c$ can be calculated directly by applying global average pooling (GAP, see Chapter 4, page 107) on the feature maps $A^k$, although it is imposed that the activation output (softmax or sigmoid function) is applied directly after the GAP layer, as suggested by the original CAM method [442]. However, this method required changes to the model's architecture, which in turn required retraining. To work around this limitation, gradient back propagation methods can also be used to solve for $w_k^c$ [348, 74]. The back propagation method does not require architecture modification or model re-training and is directly applicable to the pretrained network. The formulation based on the positive gradient, as proposed by [74] in the GradCAM++ method showed higher attention accuracy compared to the previous GradCAM method [348].

## Closing Remarks

This section has provided an overview of some of the key issues in outcome prediction, as well as reviews of the state-of-the-art in three exemplar areas. We emphasise that research into outcome prediction is not limited to these areas. Indeed, some of the most high profile work has come in other applications. Of particular note, [47] demonstrated how survival prediction could be performed in pulmonary hypertension patients using only motion estimated from cine CMR data. Such techniques, as well as those reviewed in this section, if streamlined and translated into clinical practice, could have a

major impact on risk stratification and patient management in a wide range of applications.

Next, after some self-assessment exercises, we proceed to a practical tutorial on outcome prediction, in which you will have the chance to develop machine learning models for predicting the outcome of subjects based on their cardiac shape.

## Exercises

**Exercise 1.**
What does a Kaplan-Meier curve illustrate? How could you use Kaplan-Meier curves to evaluate the effect of a treatment or intervention? Could this approach be applied to evaluate the prediction made by an AI model for outcome prediction?

**Exercise 2.**
What types of data sources can typically be exploited in outcome prediction? How does your answer change when considering traditional and AI-based approaches?

**Exercise 3.**
Explain the potential advantages/disadvantages of supervised and unsupervised analysis of data for outcome prediction.

**Exercise 4.**
As well as the three exemplar applications presented in this book, what other applications have been studied in terms of the use of AI for outcome prediction in cardiology? You may wish to perform a brief literature review to help you answer.

## Tutorial - Outcome Prediction

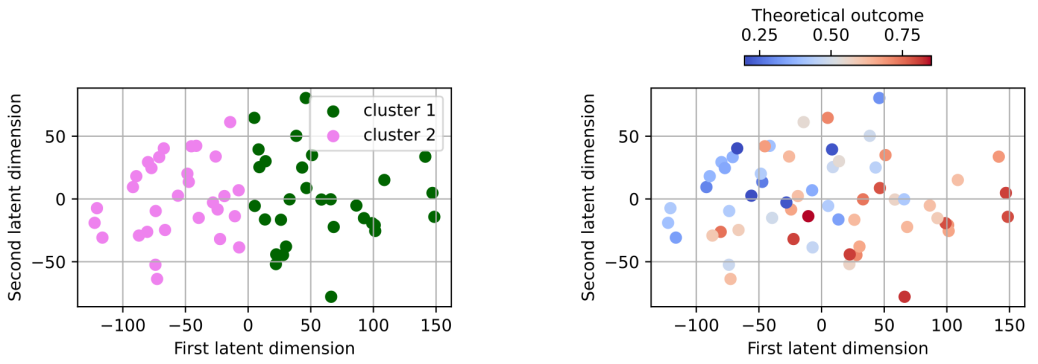
**Tutorial 5.**

As for the other notebooks, the contents of this notebook are accessible as Electronic Supplementary Material.

### Overview

In this hands-on tutorial, we aim at predicting the outcome of subjects based on their cardiac shape (here, mimicking the 2-D LV myocardial contour extracted from 4-chamber echocardiographic views). We designed synthetic data (both 2-D shapes and outcome labels) specifically tailored for the purpose of this tutorial.

You will focus on two strategies. First, (supervised) regression with the Partial Least-Squares method, which also performs dimensionality reduction: we will exploit this for interpretation purposes. Then, an unsupervised approach that chains two standard algorithms for dimensionality reduction and clustering (as shown in the figure below), which you will compare to the supervised approach.




### Objectives

- *Conduct a simple outcome prediction problem using high-dimensional data as input, with the help of the* `sckit-learn` *tools.*

- *Understand the differences between supervised and unsupervised ways of handling this problem.*

**Computing Requirements**

As for the other hands-on tutorials, this notebook starts with a brief "System setting" section, which imports the necessary packages, installs the potentially missing ones, and imports our own modules.

## Opinion

In the wider health domain deep learning has achieved successes in forecasting survival from high dimensional inputs such as cancer genomic profiles and gene expression data [430, 80] and in formulating personalized treatment recommendations [191]. Integrative approaches to risk classification have used unsupervised clustering of broad clinical variables to identify heart failure patients with distinct risk profiles [15, 353] while supervised machine learning algorithms can diagnose, risk stratify and predict adverse events from health record and registry data [34, 392, 27].

However, in an era of machine learning and AI, it is increasingly desirable that we extract quantitative biomarkers from medical images that inform on disease detection, characterization, monitoring and assessment of response to treatment. Quantitation has the potential to provide objective decision support tools in the management pathway of patients. Despite this, the quantitative potential of imaging remains under-exploited because of variability of the measurement, lack of harmonized systems for data acquisition and analysis, and crucially, a paucity of evidence on how such quantitation potentially affects clinical decision making and patient outcome. The benefit of machine learning in primary or secondary care treatment will not have an impact until a consensus is reached on how algorithmic approaches shape guideline-driven management in specific conditions and settings. Common pitfalls that can undermine machine learning-based applications include issues of transparency, reproducibility, ethics, and effectiveness [401]; and there is a pressing need for strategies to address the risk of bias when reporting performance [239].

A key challenge remains access to high quality data at scale that reflects temporal disease dynamics, heterogeneity across diverse populations, and response to interventions. Trusted research environments (TREs) facilitate accredited large scale access to health data held to common data standards which are enabling a "National Grid" of federated learning resources for researchers [160]. Such initiatives are already showing agility in providing a population-wide resource to support research on COVID-19 and cardiovascular disease [414], heralding a future where national or trans-national person-level data are discoverable and accessible to researchers through a single gateway providing a transformative substrate for outcome analysis.

The nature of what we consider health data is also being reframed enabling inferences on cardiovascular disease to be made from diverse sources such as facial imaging [233], social media activity [362] and smart wearable devices [41]. There is currently a lack of device standardization and validity testing but such approaches could offer minimally intrusive approaches for continuous monitoring of population-level trends and individual-level events. This also invites us to re-evaluate the choice of outcomes we use for risk stratification and study endpoints. Relatively few studies have comprehensively examined how lifestyle interventions may improve life expectancy free from the major diseases such as diabetes, cardiovascular disease, and cancer [229]. While a focus of current machine learning research is on disease classification or mortality prediction a key contribution to real world practice may be predicting how interventions improve "health-span" to avoid or delay the onset of multimorbidity.

machine learning itself also offers an alternative to the challenge of personalization in the context of interventional trials. More flexible data-driven approaches to classic randomized control trials may learn the relationships between the actions, context, and outcomes allowing an estimation of causal effects from the probability of receiving a treatment conditional on patient characteristics [188]. However, the goal of digitally-enabled "personalized" medical care faces serious challenges, many of which cannot be addressed through algorithmic complexity alone [410]. To learn a causal effect, we need to estimate not just the most likely outcome in a classical prediction task, but what would have happened if things had been different - a counterfactual prediction [162]. Endeavours in causal inference and causal discovery are so far largely unexplored – especially for medical imaging data. In this context, they could lead to the discovery of new applications for personalized counterfactual predictions such as what would cardiovascular function have looked like if the patient had not been exposed to a specific risk factor [65]? While conventional machine learning approaches identify risk factors associated with a future endpoint, reframing this as a counterfactual inference task improves performance where there are multiple possible causes for an outcome [327].

How might these advances in AI reshape the delivery of healthcare? Firstly, this could disrupt the conventional linear pathway of self-referral to primary care, specialist referral, and investigations eventually leading to a therapeutic intervention. Care could be more pro-active and anticipatory, integrating

data from multiple sources in the community to guide lifestyle interventions and primary prevention strategies. While traditional investigations are performed at specialist centres, AI could democratize this workflow by providing expert-level diagnostics at the point of care by physicians or even through direct-to-consumer technology. The integration of diverse data sources with innovative risk modelling could realise the 'Digital Twin' ambition of an individual-level casual framework for precision cardiology [93]. Finally, conventional diagnostic labels could become an irrelevance as we better understand the high-dimensional space that characterizes dynamic disease processes, their associated risks and effect of time-dependent interventions. This foresees healthcare providers trading discrete diagnostic classifications for improved patient-valued outcomes.

## Acknowledgements

This work was supported by the French Government, through the National Research Agency (ANR) Investments in the Future with 3IA Côte d'Azur (ANR-19-P3IA-0002) and IHU Liryc (ANR- 10-IAHU-04), and through Université Côte d'Azur STIC Doctoral School.

ND was supported by the French ANR (LABEX PRIMES of Univ. Lyon [ANR-11-LABX-0063] within the program "Investissements d'Avenir" [ANR-11-IDEX-0007], and the JCJC project "MIC-MAC" [ANR-19-CE45-0005]).