nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-019-0019-2

Deep-learning cardiac motion analysis for human survival prediction

Ghalib A. Bello[1,8], Timothy J. W. Dawes[1,2,8], Jinming Duan[1,3], Carlo Biffi[1,3], Antonio de Marvao[1], Luke S. G. E. Howard[4], J. Simon R. Gibbs[2,4], Martin R. Wilkins[5], Stuart A. Cook[1,2,6,7], Daniel Rueckert[3] and Declan P. O'Regan[1*]

Motion analysis is used in computer vision to understand the behaviour of moving objects in sequences of images. Optimizing the interpretation of dynamic biological systems requires accurate and precise motion tracking as well as efficient representations of high-dimensional motion trajectories so that these can be used for prediction tasks. Here we use image sequences of the heart, acquired using cardiac magnetic resonance imaging, to create time-resolved three-dimensional segmentations using a fully convolutional network trained on anatomical shape priors. This dense motion model formed the input to a supervised denoising autoencoder (4Dsurvival), which is a hybrid network consisting of an autoencoder that learns a task-specific latent code representation trained on observed outcome data, yielding a latent representation optimized for survival prediction. To handle right-censored survival outcomes, our network used a Cox partial likelihood loss function. In a study of 302 patients, the predictive accuracy (quantified by Harrell's *C*-index) was significantly higher ($P = 0.0012$) for our model $C = 0.75$ (95% CI: 0.70–0.79) than the human benchmark of $C = 0.59$ (95% CI: 0.53–0.65). This work demonstrates how a complex computer vision task using high-dimensional medical image data can efficiently predict human survival.

Techniques for vision-based motion analysis aim to understand the behaviour of moving objects in image sequences[1]. In this domain, deep-learning architectures have achieved a wide range of competencies for object tracking, action recognition and semantic segmentation[2]. Making predictions about future events from the current state of a moving three-dimensional (3D) scene depends on learning correspondences between patterns of motion and subsequent outcomes. Such relationships are important in biological systems that exhibit complex spatio-temporal behaviour in response to stimuli or as a consequence of disease processes. Here we use recent advances in machine learning for visual processing tasks to develop a generalizable approach for modelling time-to-event outcomes from time-resolved 3D sensory input. We tested this on the challenging task of predicting survival due to heart disease through analysis of cardiac imaging.

The motion dynamics of the beating heart are a complex rhythmic pattern of nonlinear trajectories regulated by molecular, electrical and biophysical processes[3]. Heart failure is a disturbance of this coordinated activity characterized by adaptations in cardiac geometry and motion that lead to impaired organ perfusion[4]. For this prediction task, we studied patients diagnosed with pulmonary hypertension, characterized by right ventricular (RV) dysfunction, as this is a disease with high mortality where the choice of treatment depends on individual risk stratification[5]. Our input data were derived from cardiac magnetic resonance (CMR), which acquires imaging of the heart in any anatomical plane for dynamic assessment of function. While explicit measurements of performance obtained from myocardial motion tracking detect early contractile dysfunction and act as discriminators of different pathologies[6,7], we hypothesized that learned features of complex 3D cardiac motion would provide enhanced prognostic accuracy.

A major challenge for medical image analysis has been to automatically derive quantitative and clinically relevant information in patients with disease phenotypes. Our method employs a fully convolutional network (FCN) to learn a cardiac segmentation task from manually labelled priors. The outputs are smooth 3D renderings of frame-wise cardiac motion that are used as input data to a supervised denoising autoencoder (DAE) prediction network that we refer to as 4Dsurvival. The aim is to learn latent representations robust to noise and salient for survival prediction. We then compared our model to a benchmark of conventional human-derived volumetric indices and clinical risk factors in survival prediction.

Results

Baseline characteristics. Data from all 302 patients with incident pulmonary hypertension were included for analysis. Objective diagnosis was made according to haemodynamic and imaging criteria[5]. Patients were investigated between 2004 and 2017, and were followed-up until 27 November 2017 (median 371 days). All-cause mortality was 28% (85 of 302). Table 1 summarizes characteristics of the study sample at the date of diagnosis. No subjects' data were excluded.

Magnetic resonance image processing. Automatic segmentation of the ventricles from gated CMR images was performed for each slice position at each of 20 temporal phases, producing a total of 69,820 label maps for the cohort (Fig. 1a). Image registration was used to track the motion of corresponding anatomic points. Data for each subject were aligned, producing a dense model of cardiac

[1]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [2]National Heart and Lung Institute, Imperial College London, London, UK. [3]Department of Computing, Imperial College London, London, UK. [4]Imperial College Healthcare NHS Trust, London, UK. [5]Division of Experimental Medicine, Department of Medicine, Imperial College London, London, UK. [6]National Heart Centre Singapore, Singapore, Singapore. [7]Duke-NUS Graduate Medical School, Singapore, Singapore. [8]These authors contributed equally: Ghalib A. Bello, Timothy J. W. Dawes. *e-mail: declan.oregan@imperial.ac.uk

Table 1 | Patient characteristics at the baseline (date of MRI scan)

Characteristic	*n*	% or mean ± s.d.
Age (years)		62.9 ± 14.5
Body surface area (m^2)		1.92 ± 0.25
Male	169	56
Race		
Caucasian	215	71.2
Asian	7	2.3
Black	13	4.3
Other	28	9.3
Unknown	39	12.9
World Health Organization functional class		
I	1	0
II	45	15
III	214	71
IV	42	14
Haemodynamics		
Systolic blood pressure (mmHg)		131.5 ± 25.2
Diastolic blood pressure (mmHg)		75 ± 13
Heart rate (beats min^{-1})		69.8 ± 22.5
Mean right atrial pressure (mmHg)		9.9 ± 5.8
Mean pulmonary artery pressure (mmHg)		44.1 ± 12.6
Pulmonary vascular resistance (Wood units)		8.9 ± 5.0
Cardiac output (l min^{-1})		4.3 ± 1.5
LV volumetry		
LV ejection fraction (%)		61 ± 11.1
LV end-diastolic volume (ml)		110 ± 37.4
LV end-systolic volume (ml)		44 ± 22.9
RV volumetry		
RV ejection fraction (%)		38 ± 13.7
RV end-diastolic volume (ml)		194 ± 62
RV end-systolic volume (ml)		125 ± 59.3
RV strain		
Longitudinal (%)		−16.8 ± 4.7
Radial (%)		+18.0 ± 4.4
Circumferential (%)		−9.6 ± 7.0

motion across the patient population (Fig. 1b) that was then used as an input to the 4Dsurvival network.

Predictive performance. Bootstrapped internal validation was applied to the 4Dsurvival and benchmark models. The apparent predictive accuracy for 4Dsurvival was $C=0.86$ and the optimism-corrected value was $C=0.75$ (95% confidence interval (CI): 0.70–0.79). The 4Dsurvival model outperformed (1) benchmark models of volumetric CMR parameters ($P=0.0012$): apparent predictive accuracy $C=0.60$ and optimism-adjusted $C=0.59$ (95% CI: 0.53–0.65); (2) myocardial strain parameters ($P=0.016$): apparent predictive accuracy $C=0.64$ and optimism-adjusted $C=0.61$ (95% CI: 0.57–0.66); and (3) a joint analysis of both imaging and clinical risk factors ($P=0.006$): apparent predictive accuracy $C=0.66$ and optimism-adjusted $C=0.64$ (95% CI: 0.57–0.70). Figure 2 shows Kaplan–Meier plots that depict the survival probability estimates over time, stratified by risk groups defined by each model's predictions (see Supplementary Information for details). After bootstrap validation, a final model was created using the training and optimization procedure outlined in the Methods (optimal hyperparameters for this model are summarized in Table 2).

Visualization of learned representations. To assess the ability of the 4Dsurvival network to learn discriminative features from the data, we examined the encoded representations by projection to 2D space using Laplacian eigenmaps[8] (Fig. 3a). In this figure, each subject is represented by a point, the colour of which is based on the subject's survival time (that is, the time elapsed from the baseline (date of magnetic resonance imaging (MRI) scan) to death (for uncensored patients), or to the most recent follow-up date (for censored patients)). Survival time was truncated at 7 years for ease of visualization. As is evident from the plot, our network's compressed representations of 3D motion input data show distinct patterns of clustering according to survival time. Figure 3a also shows visualizations of RV motion for two exemplar subjects at opposite ends of the risk spectrum. We also assessed the extent to which motion in various regions of the RV contributed to overall survival prediction. Fitting univariate linear models to each vertex in the mesh (see Methods for full details), we computed the association between the magnitude of cardiac motion and the 4Dsurvival network's predicted risk score, yielding a set of regression coefficients (one per vertex) that were then mapped onto a template RV mesh, producing a 3D saliency map (Fig. 3b). These show the contribution from spatially distant but functionally synergistic regions of the RV in influencing survival in pulmonary hypertension.

Discussion

Machine-learning algorithms have been used in a variety of motion analysis tasks from classifying complex traits to predicting future events from a given scene[9–11]. We show that compressed representations of a dynamic biological system moving in 3D space offer a powerful approach for time-to-event analysis. In this example, we demonstrate the effectiveness of a deep-learning algorithm, trained to find correspondences between heart motion and patient outcomes, for efficiently predicting human survival.

The traditional paradigm of epidemiological research is to 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[12]. 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[13]. Integrative approaches to risk classification have used unsupervised clustering of broad clinical variables to identify heart failure patients with distinct risk profiles[14,15], while supervised machine-learning algorithms can diagnose, risk stratify and predict adverse events from health record and registry data[16–18]. 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[19,20], and in formulating personalized treatment recommendations[21].

With the exception of natural image tasks, such as classification of skin lesions[22], biomedical imaging poses a number of challenges for machine learning as the datasets are often of limited scale, inconsistently annotated and typically high-dimensional[23]. Architectures predominantly based on convolutional neural nets, often using data augmentation strategies, have been successfully applied in computer vision tasks to enhance clinical images, segment organs and classify lesions[24,25]. Segmentation of cardiac images in the time domain is a well-established visual correspondence task that has recently achieved expert-level performance with FCN architectures[26]. Atlas-based analyses of cardiac geometry have demonstrated their value in disease classification and visualization[27–29]. Supervised principal component analysis of semi-automated segmentations has

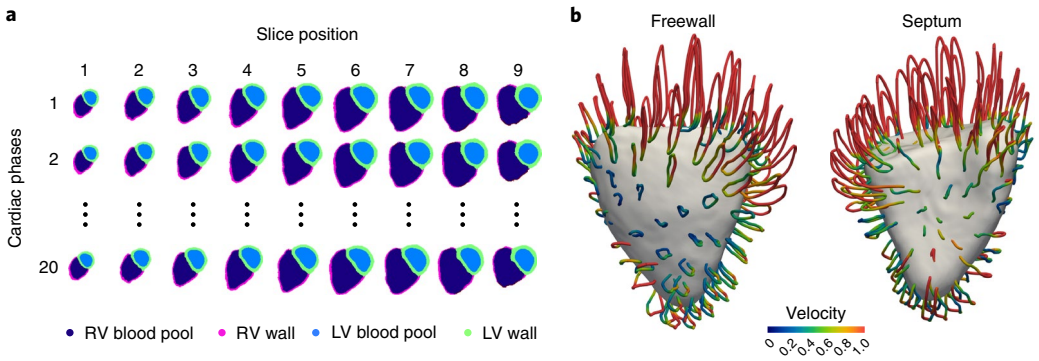

Fig. 1 | Segmentation and motion estimation. a, An example of an automatic cardiac image segmentation of each short-axis cine image from the apex (slice 1) to the base (slice 9) across 20 temporal phases. Data were aligned to a common reference space to build a population model of cardiac motion. **b**, Trajectory of RV contraction and relaxation averaged across the study population plotted as looped pathlines for a subsample of 100 points on the heart (magnification factor of ×4). The colour represents the relative myocardial velocity at each phase of the cardiac cycle. A surface-shaded model of the heart is shown at end-systole. These dense myocardial motion fields for each patient were used as an input to the prediction network.

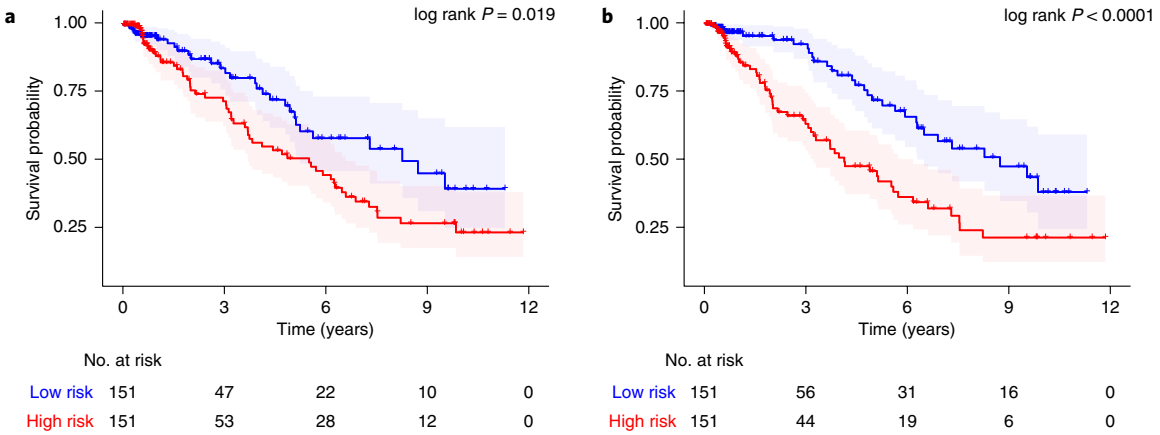

Fig. 2 | Kaplan–Meier Plots. a,**b**, Kaplan–Meier plots for a conventional parameter model using a composite of manually derived volumetric measures (**a**), and a deep-learning prediction model (4Dsurvival) whose input was time-resolved 3D models of cardiac motion (**b**). For both models, patients were divided into low- and high-risk groups by median risk score. Survival function estimates for each group (with 95% CI) are shown. For each plot, the logrank test was performed to compare survival curves between risk groups (conventional parameter model: $\chi^2 = 5.5$, $P = 0.019$; 4Dsurvival: $\chi^2 = 15.6$, $P < 0.0001$).

Table 2 | Hyperparameter search ranges for the deep-learning network (second column) and optimum hyperparameter values in the final model (third column)

Hyperparameter	Search range	Optimized value
Dropout	[0.1, 0.9]	0.71
Number of nodes in hidden layers	[75, 250]	78
Latent code dimensionality (h)	[5, 20]	13
Reconstruction loss penalty (α)	[0.3, 0.7]	0.6
Learning rate	$[10^{-6}, 10^{-4.5}]$	$10^{-4.86}$
L1 regularization penalty	$[10^{-7}, 10^{-4}]$	$10^{-5.65}$

shown prognostic utility compared to conventional parameters[30], but requires human selection of anatomical features and relies on simple predefined motion characteristics. In this work, we harness the power of deep learning for both automated image analysis and inference—learning features predictive of survival from 3D cardiac motion using nonlinear data transformations.

Autoencoding is a dimensionality reduction technique in which an encoder takes an input and maps it to a latent representation (lower-dimensional space) that is in turn mapped back to the space of the original input. The last step represents an attempt to 'reconstruct' the input from the compressed (latent) representation, and this is performed in such a way as to minimize the reconstruction error (that is, the degree of discrepancy between the input and its reconstructed version). Our algorithm is based on a DAE, a type of autoencoder that aims to extract more robust latent representations by corrupting the input with stochastic noise[31]. While conventional autoencoders are used for unsupervised learning tasks we extend recent proposals for supervised autoencoders in which the learned representations are both reconstructive and discriminative[32–38]. We achieved this by adding a prediction branch to the network with a loss function for survival inspired by the Cox proportional hazards model. A hybrid loss function, optimizing the trade-off between survival prediction and accurate input reconstruction, is calibrated during training. The compressed representations of 3D motion predict survival more accurately than a composite measure of conventional manually derived parameters measured on the same images and the improvement in performance is independent of clinical risk factors.

The main limitation of our study is relying on internal validation to evaluate predictive performance, and so the next step towards implementation is to train on larger and more diverse multicentre patient groups using image data and other prognostic variables, before performing external validation of survival prediction in a clini-

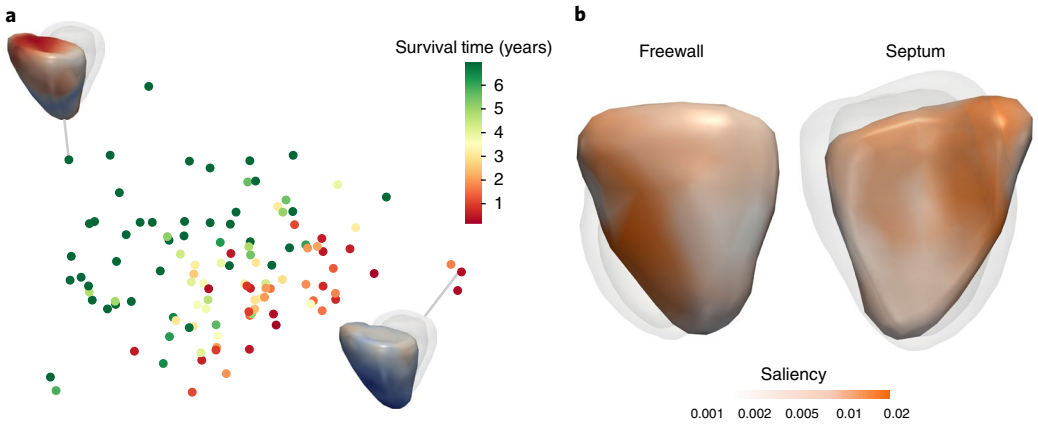

Fig. 3 | Model interpretation. a, A 2D projection of latent representations of cardiac motion in the 4Dsurvival network labelled by survival time. A visualization of RV motion is shown for two patients with contrasting risks. **b**, A saliency map showing regional contributions to survival prediction by RV motion. Absolute regression coefficients are expressed on a log scale.

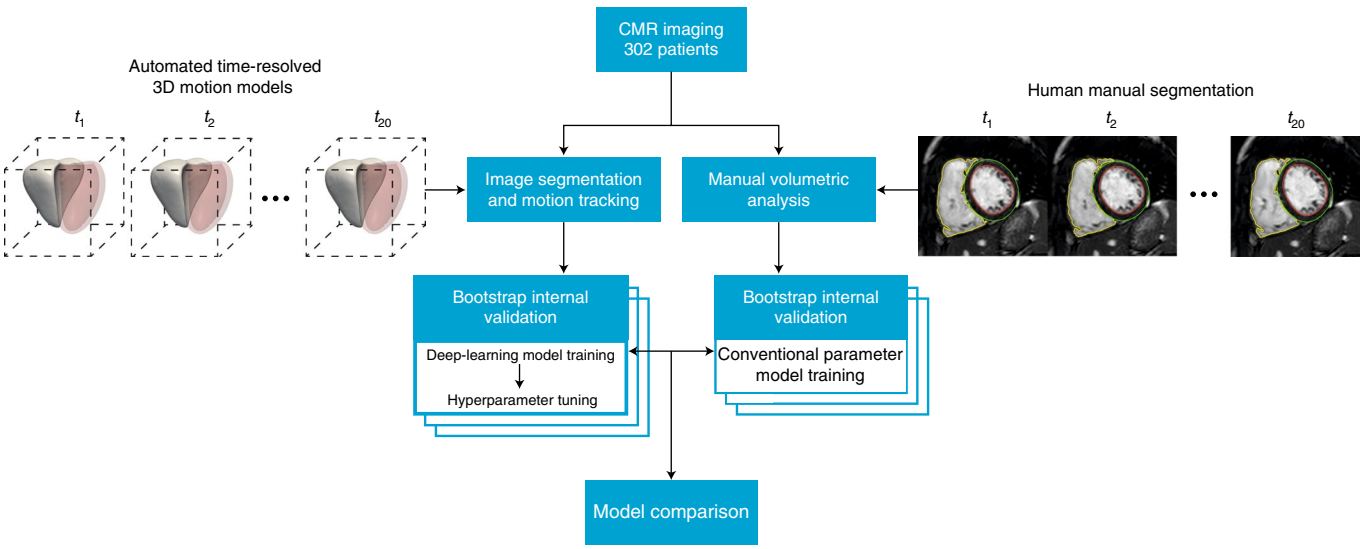

Fig. 4 | Flow chart showing the design of the study. In total, 302 patients with CMR imaging had both manual volumetric analysis and automated image segmentation (right ventricle shown in solid white, left ventricle in red) across 20 temporal phases ($t=1, ..., 20$). The internal validity of the predictive performance of a conventional parameter model and a deep-learning motion model was assessed using bootstrapping.

cal setting against a benchmark of established risk prediction scores[39]. Autoencoders may be more prone to over-fitting than methods such as principal component analysis and are more computationally expensive to train. We mitigated over-fitting using dropout and L1 regularization, and reduced the input space by down-sampling spatially correlated data. We used routinely acquired clinical data and applied normalization to compare motion acquired at different temporal resolutions. Improvement in performance may be achievable at higher temporal resolutions, but would also increase the dimension of the input data. CMR provides accurate assessment of cardiac function but other imaging modalities may offer complementary prognostic markers[40]. Further enhancement in predictive performance may be achievable by modelling multiple observations over time, for instance using long short-term memory and other recurrent neural network architectures[41,42], and handling independent competing risks[43].

Our approach enables fully automated and interpretable predictions of survival from moving clinical images—a task that has not been previously achieved in heart failure or other disease domains. This fast and scalable method is readily deployable and could have a substantial impact on clinical decision-making and the understanding of disease mechanisms. Extending this approach to other conditions where motion is predictive of survival is constrained only by the availability of suitable training cases with known outcomes.

Image acquisition and computational methods

In the following sections we describe patient data collection, the CMR protocol for image acquisition, our FCN network for image segmentation and construction of 3D cardiac motion models.

Study population. In a single-centre observational study, we analysed data collected from patients referred to the National Pulmonary Hypertension Service at the Imperial College Healthcare NHS Trust between May 2004 and October 2017. The study was approved by the Heath Research Authority and all participants gave written informed consent. Criteria for inclusion were a documented diagnosis of Group 4 pulmonary hypertension investigated by right heart catheterization with a mean pulmonary artery pressure ≥25 mmHg and pulmonary capillary wedge pressure <15 mmHg; and signs of chronic

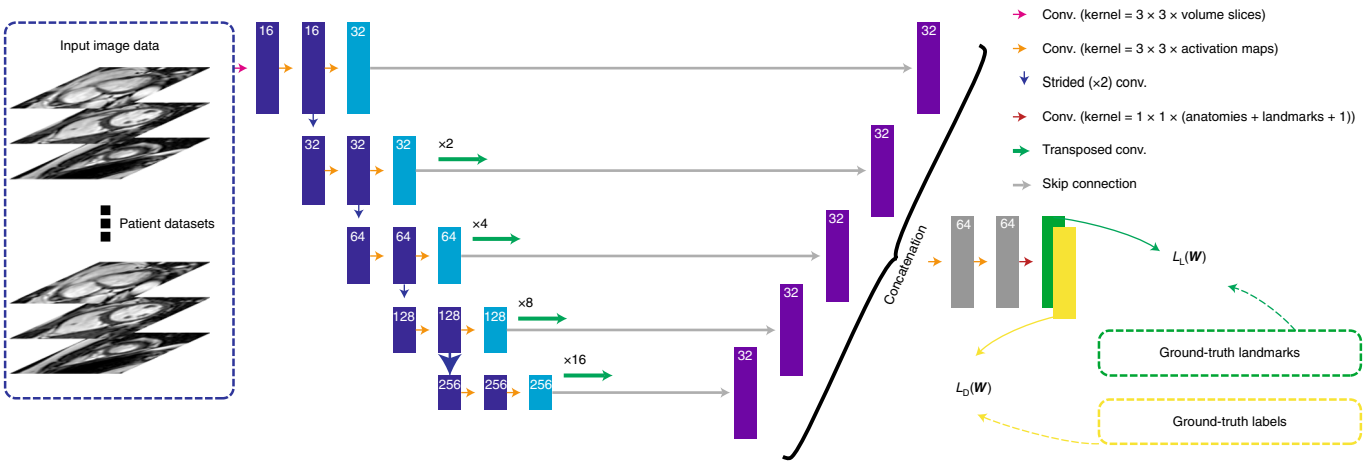

Fig. 5 | The architecture of the segmentation algorithm. A FCN takes each stack of cine images as an input, applies a branch of convolutions, learns image features from fine to coarse levels, concatenates multi-scale features and finally predicts the segmentation and landmark location probability maps simultaneously. These maps, together with the ground-truth landmark locations and label maps, are then used in the loss function (see equation (1)) that is minimized via stochastic gradient descent.

thrombo-embolic disease present on either ventilation–perfusion scintigraphy or computed tomography pulmonary angiography[44]. All patients were treated in accordance with current guidelines including medical and surgical therapy as clinically indicated[5].

MRI image acquisition, processing and computational image analysis. The CMR protocol has been previously described in detail[30]. Briefly, imaging was performed on a 1.5T Achieva (Philips), using a standard clinical protocol based on international guidelines[45]. The specific images analysed in this study were retrospectively gated cine sequences, in the short-axis plane of the heart, with a reconstructed spatial resolution of $1.3\times1.3\times10.0$ mm and a typical temporal resolution of 29 ms. Images were stored on an open-source data management system[46]. Manual volumetric analysis of the images (Fig. 4) was independently performed by accredited physicians using proprietary software (cmr42, Circle Cardiovascular Imaging) according to international guidelines with access to all available images for each subject and no analysis time constraint[47]. The derived parameters included the strongest and most well-established volumetric and functional CMR findings for prognostication reported in disease-specific meta-analyses[48,49].

We developed a convolutional neural net combined with image registration for shape-based biventricular segmentation of the CMR images. The pipeline method has three main components: segmentation, landmark localization and shape registration. First, a 2.5D multi-task FCN is trained to effectively and simultaneously learn segmentation maps and landmark locations from manually labelled volumetric CMR images. Second, multiple high-resolution 3D atlas shapes are propagated onto the network segmentation to form a smooth segmentation model. This step effectively induces a hard anatomical shape constraint and is fully automatic due to the use of predicted landmarks from the network.

We treat the problem of predicting segmentations and landmark locations as a multi-task classification problem. First, let us formulate the learning problem as follows: we denote the input training dataset by $S=\{(U_i, R_i, L_i), i=1, \ldots, N_t\}$, where N_t is the sample size of the training data, $U_i=\{u_j^i, j=1, \ldots, |U_i|\}$ is the raw input CMR volume, $R_i=\{r_j^i, j=1, \ldots, |R_i|\}$ and $r_j^i\in\{1, \ldots, N_r\}$ are the ground-truth region labels for volume U_i ($N_r=5$ representing 4 regions and background) and $L_i=\{l_j^i, j=1, \ldots, |L_i|\}$ and $l_j^i\in\{1, \ldots, N_l\}$ are the labels representing ground-truth landmark locations for U_i ($N_l=7$ representing 6 landmark locations and background). Note that $|U_i|=|R_i|=|L_i|$ stands for the total number of voxels in a CMR volume. Let $\mathbf{W}$ denote the set of all network layer parameters. In a supervised setting, we minimize the following objective function via standard (backpropagation) stochastic gradient descent:

$$L(\mathbf{W})=L_D(\mathbf{W})+\alpha L_L(\mathbf{W})+\beta\ \|\mathbf{W}\|_F^2 \quad (1)$$

where α and β are weight coefficients balancing the corresponding terms. $L_D(\mathbf{W})$ is the region-associated loss that enables the network to predict segmentation maps. $L_L(\mathbf{W})$ is the landmark-associated loss for predicting landmark locations. $\|\mathbf{W}\|_F^2$, known as the weight decay term, represents the Frobenius norm on the weights $\mathbf{W}$. This term is used to prevent the network from overfitting. The training problem is therefore to estimate the parameters $\mathbf{W}$ associated with all of the convolutional layers. By minimizing equation (1), the network is able to simultaneously predict segmentation maps and landmark locations. The definitions of the loss functions $L_D(\mathbf{W})$ and $L_L(\mathbf{W})$, used for predicting landmarks and segmentation labels, have been described previously[50].

The FCN segmentations are used to perform a non-rigid registration using cardiac atlases built from >1,000 high-resolution images[51], allowing shape constraints to be inferred. This approach produces accurate, high-resolution and anatomically smooth segmentation results from input images with low through-slice resolution, thus preserving clinically important global anatomical features. The data were split in the ratio 70:30 for training and evaluation, respectively. Motion tracking was performed for each subject using a 4D spatio-temporal B-spline image registration method with a sparseness regularization term[52]. The motion field estimate is represented by a displacement vector at each voxel and at each time frame $t=1$, ..., 20. Temporal normalization was performed before motion estimation to ensure consistency across the cardiac cycle.

Spatial normalization of each patient's data was achieved by registering the motion fields to a template space. A template image was built by registering the high-resolution atlases at the end-diastolic frame and then computing an average intensity image. In addition, the corresponding ground-truth segmentations for these high-resolution images were averaged to form a segmentation of the template image. A template surface mesh was then reconstructed from its segmentation using a 3D surface reconstruction algorithm. The motion field estimate lies within the reference space of each subject and so to enable inter-subject comparison all the segmentations were aligned to this template space by non-rigid B-spline image registration[53]. We then warped the template mesh using the resulting

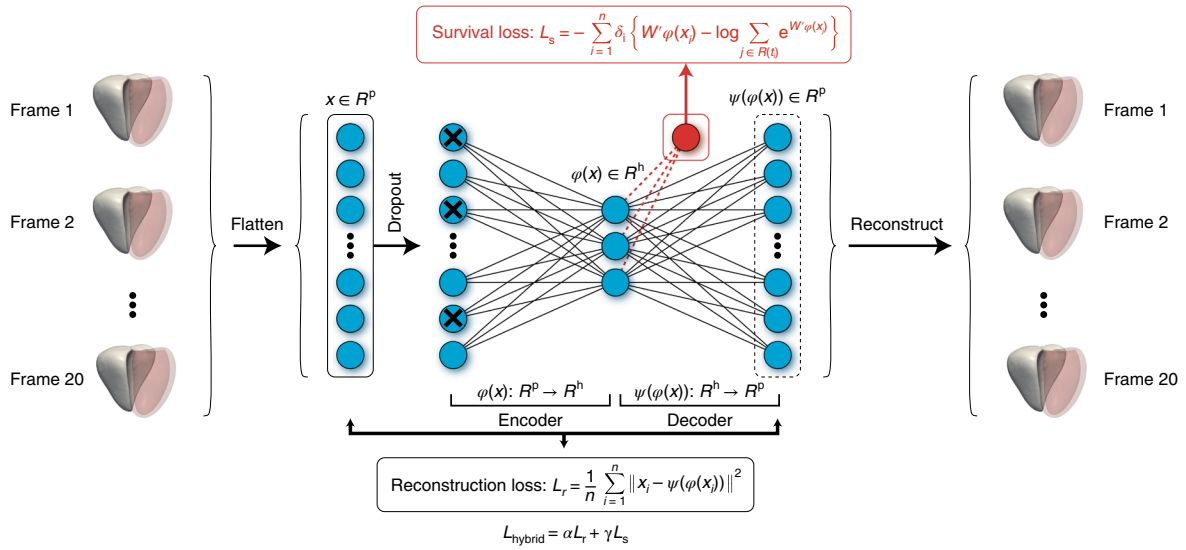

Fig. 6 | The architecture of the prediction network. The prediction network (4Dsurvival) is a DAE that takes time-resolved cardiac motion meshes as its input (right ventricle shown in solid white, left ventricle in red). For the sake of simplicity, two hidden layers, one immediately preceding and the other immediately following the central layer (latent code layer), have been excluded from the diagram. The autoencoder learns a task-specific latent code representation trained on observed outcome data, yielding a latent representation optimized for survival prediction that is robust to noise. The actual number of latent factors is treated as an optimizable parameter.

non-rigid deformation and mapped it back to the template space. Twenty surface meshes, one for each temporal frame, were subsequently generated by applying the estimated motion fields to the warped template mesh accordingly. Consequently, the surface mesh of each subject at each frame contained the same number of vertices (18,028), which maintained their anatomical correspondence across temporal frames, and across subjects (Fig. 5).

Characterization of RV motion. The time-resolved 3D meshes described in the previous section were used to produce a relevant representation of cardiac motion—in this example of right-side heart failure limited to the RV. For this purpose, we utilized a sparser version of the meshes (down-sampled by a factor of ~90) with 202 vertices. Anatomical correspondence was preserved in this process by utilizing the same vertices across all meshes. To characterize motion, we adapted an approach outlined in Bai et al.[54].

This approach is used to produce a simple numerical representation of the trajectory of each vertex (that is, the path each vertex traces through space during a cardiac cycle (see Fig. 1b)). Let (x_{vt}, y_{vt}, z_{vt}) represent the Cartesian coordinates of vertex v ($v = 1, ..., 202$) at the tth time frame ($t = 1, ..., 20$) of the cardiac cycle. At each time frame $t = 2, 3, \ldots, 20$, we compute the coordinate-wise displacement of each vertex from its position at time frame 1. This yields the following 1D input vector:

$$\mathbf{x} = (x_{vt} - x_{v1},\quad y_{vt} - y_{v1},\quad z_{vt} - z_{v1})_{2 \le t \le 20}^{1 \le v \le 202} \tag{2}$$

Vector $\mathbf{x}$ has length 11,514 ($3 \times 19 \times 202$), and was used as the input feature for our prediction network.

4Dsurvival network model

Our 4Dsurvival network structure is summarized in Fig. 6.

Network design and training. We aimed to produce an architecture capable of learning a low-dimensional representation of RV motion that robustly captures prognostic features indicative of poor survival. The architecture's hybrid design combines a DAE[55], with a Cox proportional hazards model (described below)[56].

As before, we denote our input vector by $\mathbf{x} \in \mathbb{R}^{d_p}$, where $d_p = 11{,}514$, the input dimensionality. Our network is based on a DAE, an autoencoder variant that learns features robust to noise[55]. The input vector $\mathbf{x}$ feeds directly into a stochastic masking filter layer that produces a corrupted version of $\mathbf{x}$. The masking is implemented using random dropout[57]; that is, we randomly set a fraction m of the elements of vector $\mathbf{x}$ to zero (the value of m is treated as an optimizable network parameter). The corrupted input from the masking filter is then fed into a hidden layer, the output of which is in turn fed into a central layer. This central layer represents the latent code (that is, the encoded/compressed representation of the input). This central layer is referred to as the 'code', or 'bottleneck' layer. Therefore, we may consider the encoder as a function $\phi(\cdot)$ mapping the input $\mathbf{x} \in \mathbb{R}^{d_p}$ to a latent code $\phi(\mathbf{x}) \in \mathbb{R}^{d_h}$, where $d_h \ll d_p$ (for notational convenience, we consider the corruption step as part of the encoder). This produces a compressed representation whose dimensionality is much lower than that of the input (an undercomplete representation)[58]. Note that the number of units in the encoder's hidden layer, and the dimensionality of the latent code (d_h) are not predetermined but, rather, treated as optimizable network parameters. The latent code $\phi(\mathbf{x})$ is then fed into the second component of the DAE, a multilayer decoder network that upsamples the code back to the original input dimension d_p. Like the encoder, the decoder has one intermediate hidden layer that feeds into the final layer, which in turn outputs a decoded representation (with dimension d_p matching that of the input). The size of the decoder's intermediate hidden layer is constrained to match that of the encoder network, to give the autoencoder a symmetric architecture. Dissimilarity between the original (uncorrupted) input $\mathbf{x}$ and the decoder's reconstructed version (denoted here by $\psi(\phi(\mathbf{x}))$) is penalized by minimizing a loss function of general form $L(\mathbf{x}, \psi(\phi(\mathbf{x})))$. Herein, we chose a simple mean squared error form for L:

$$L_r = \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{x}_i - \psi(\phi(\mathbf{x}_i)) \|^2 \tag{3}$$

where n represents the sample size. Minimizing this loss forces the autoencoder to reconstruct the input from a corrupted/incomplete

version, thereby facilitating the generation of a latent representation with robust features. Further, to ensure that these learned features are actually relevant for survival prediction, we augmented the auto-encoder network by adding a prediction branch. The latent representation learned by the encoder $\phi(\mathbf{x})$ is therefore linked to a linear predictor of survival (see equation (4) below), in addition to the decoder. This encourages the latent representation $\phi(\mathbf{x})$ to contain features that are simultaneously robust to noisy input and salient for survival prediction. The prediction branch of the network is trained with observed outcome data (that is, survival/follow-up time). For each subject, this is the time elapsed from MRI acquisition until death (all-cause mortality), or if the subject is still alive, the last date of follow-up. Furthermore, patients receiving surgical interventions were censored at the date of surgery. This type of outcome is called a right-censored time-to-event outcome[59], and is typically handled using survival analysis techniques, the most popular of which is Cox's proportional hazards regression model[56]:

$$\log\frac{h_i(t)}{h_0(t)} = \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_p z_{ip} \tag{4}$$

Here, $h_i(t)$ represents the hazard function for subject i; that is, the 'chance' (normalized probability) of subject i dying at time t. The term $h_0(t)$ is a baseline hazard level to which all subject-specific hazards $h_i(t)$ ($i=1, \ldots, n$) are compared. The key assumption of the Cox survival model is that the hazard ratio $h_i(t)/h_0(t)$ is constant with respect to time (proportional hazards assumption)[56]. The natural logarithm of this ratio is modelled as a weighted sum of a number of predictor variables (denoted here by $z_{i1}, \ldots, z_{ip}$), where the weights/coefficients are unknown parameters denoted by $\beta_1, \ldots, \beta_p$. These parameters are estimated via maximization of the Cox proportional hazards partial likelihood function:

$$\log\mathcal{L}(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta' \mathbf{z}_i - \log \sum_{j \in R(t_i)} \mathrm{e}^{\beta' \mathbf{z}_j} \right\} \tag{5}$$

In the expression above, $\mathbf{z}_i$ is the vector of predictor/explanatory variables for subject i, δ_i is an indicator of subject i's status (0 = alive, 1 = dead) and $R(t_i)$ represents subject i's risk set (that is, subjects still alive (and thus at risk) at the time subject i died or became censored ($\{j : t_j > t_i\}$)).

We adapt this loss function for our neural network architecture as follows:

$$L_s = -\sum_{i=1}^{n} \delta_i \left\{ \mathbf{W}' \phi(\mathbf{x}_i) - \log \sum_{j \in R(t_i)} \mathrm{e}^{\mathbf{W}' \phi(\mathbf{x}_j)} \right\} \tag{6}$$

The term $\mathbf{W}'$ denotes a $(1 \times d_h)$ vector of weights, which when multiplied by the d_h-dimensional latent code $\phi(\mathbf{x})$ yields a single scalar ($\mathbf{W}'\phi(\mathbf{x}_i)$) representing the survival prediction (specifically, natural logarithm of the hazard ratio) for subject i. Note that this makes the prediction branch of our 4Dsurvival network essentially a simple linear Cox proportional hazards model, and the predicted output may be seen as an estimate of the log hazard ratio (see equation (4)).

For our network, we combine this survival loss with the reconstruction loss from equation (3) to form a hybrid loss given by:

$$\begin{aligned} L_{\text{hybrid}} &= \alpha L_r + \gamma L_s \\ &= \alpha \left[\frac{1}{n} \sum_{i=1}^{n} \| \mathbf{x}_i - \psi(\phi(\mathbf{x}_i)) \|^2 \right] \\ &\quad + \gamma \left[-\sum_{i=1}^{n} \delta_i \left\{ \mathbf{W}' \phi(\mathbf{x}_i) - \log \sum_{j \in R(t_i)} e^{\mathbf{W}' \phi(\mathbf{x}_j)} \right\} \right] \end{aligned} \tag{7}$$

The terms α and γ are used to calibrate the contributions of each term to the overall loss (that is, to control the trade-off between survival prediction versus accurate input reconstruction). During network training, they are treated as optimizable network hyperparameters, with γ chosen to equal $1-\alpha$ for convenience.

The loss function was minimized via backpropagation. To avoid overfitting and to encourage sparsity in the encoded representation, we applied L1 regularization. The rectified linear unit activation function was used for all layers, except the output layers (linear activation was used for these layers). Using the adaptive moment estimation (Adam) algorithm, the network was trained for 100 epochs with a batch size of 16 subjects. The learning rate is treated as a hyperparameter (see Table 2). During training, the random dropout (input corruption) was repeated at every backpropagation pass. The network was implemented and trained in the Python deep-learning libraries TensorFlow[60] and Keras[61], on a high-performance computing cluster with an Intel Xeon E5-1660 CPU and NVIDIA TITAN Xp GPU. The entire training process (including hyperparameter search and bootstrap-based internal validation (see subsections below)) took a total of 131 h.

Hyperparameter tuning. To determine optimal hyperparameter values, we utilized particle swarm optimization (PSO)[62], a gradient-free meta-heuristic approach to finding optima of a given objective function. Inspired by the social foraging behaviour of birds, PSO is based on the principle of swarm intelligence, which refers to problem-solving ability that arises from the interactions of simple information-processing units[63]. In the context of hyperparameter tuning, it can be used to maximize the prediction accuracy of a model with respect to a set of potential hyperparameters[64]. We used PSO to choose the optimal set of hyperparameters from among predefined ranges of values (summarized in Table 2). We ran the PSO algorithm for 50 iterations, at each step evaluating candidate hyperparameter configurations using sixfold cross-validation. The hyperparameters at the final iteration were chosen as the optimal set. This procedure was implemented via the Python library Optunity[65].

Model validation and comparison. *Predictive accuracy metric.* Discrimination was evaluated using Harrell's concordance index[66], an extension of area under the receiver operating characteristic curve to censored time-to-event data:

$$C = \frac{\sum_{i,j} \delta_i \times I(\eta_i > \eta_j) \times I(t_i < t_j)}{\sum_{i,j} \delta_i \times I(t_i < t_j)} \tag{8}$$

In the above equation, the indices i and j refer to pairs of subjects in the sample and I denotes an indicator function that evaluates to 1 if its argument is true (and 0 otherwise). The symbols η_i and η_j denote the predicted risks for subjects i and j. The numerator tallies the number of subject pairs (i, j) where the pair member with greater predicted risk has shorter survival, representing agreement (concordance) between the model's risk predictions and ground-truth survival outcomes. Multiplication by δ_i restricts the sum to subject pairs where it is possible to determine who died first (that is, informative pairs). The C index therefore represents the fraction of informative pairs exhibiting concordance between predictions and outcomes. The index has a similar interpretation to the area under the receiver operating characteristic curve (and consequently, the same range).

Internal validation. To get a sense of how well our model would generalize to an external validation cohort, we assessed its predictive accuracy within the training sample using a bootstrap-based procedure recommended in the guidelines for transparent reporting of a multivariable model for individual prognosis or diagnosis[67]. This

procedure attempts to derive realistic, 'optimism-adjusted' estimates of the model's generalization accuracy using the training sample[68]. Below, we outline the steps of the procedure.

In step 1, a prediction model was developed on the full training sample (size n), utilizing the hyperparameter search procedure discussed above to determine the best set of hyperparameters. Using the optimal hyperparameters, a final model was trained on the full sample. Then the Harrell's concordance index (C) of this model was computed on the full sample, yielding the apparent accuracy (that is, the inflated accuracy obtained when a model is tested on the same sample on which it was trained/optimized).

In step 2, a bootstrap sample was generated by carrying out n random selections (with replacement) from the full sample. On this bootstrap sample, we developed a model (applying exactly the same training and hyperparameter search procedure used in step 1) and computed C for the bootstrap sample (henceforth referred to as bootstrap performance). Then the performance of this bootstrap-derived model on the original data (the full training sample) was also computed (henceforth referred to as test performance).

In step 3, for each bootstrap sample, the optimism was computed as the difference between the bootstrap performance and the test performance.

In step 4, steps 2 and 3 were repeated B times (where $B=100$).

In step 5, the optimism estimates derived from steps 2–4 were averaged across the B bootstrap samples and the resulting quantity was subtracted from the apparent predictive accuracy from step 1.

This procedure yields an optimism-corrected estimate of the model's concordance index:

$$C_{\text{corrected}} = C_{\text{full}}^{\text{full}} - \frac{1}{B}\sum_{b=1}^{B}\left(C_b^b - C_b^{\text{full}}\right) \quad (9)$$

Above, the symbol $C_{s_1}^{s_2}$ refers to the concordance index of a model trained on sample s_1 and tested on sample s_2. The first term refers to the apparent predictive accuracy (that is, the (inflated) concordance index obtained when a model trained on the full sample is then tested on the same sample). The second term is the average optimism (difference between bootstrap performance and test performance) over the B bootstrap samples. It has been demonstrated that this sample-based average is a nearly unbiased estimate of the expected value of the optimism that would be observed in external validation[68–71]. Subtraction of this optimism estimate from the apparent predictive accuracy gives the optimism-corrected predictive accuracy.

Conventional parameter model. As a benchmark comparison to our RV motion model, we trained a Cox proportional hazards model using conventional RV volumetric indices including RV end-diastolic volume, RV end-systolic volume and the difference between these measures expressed as a percentage of RV end-diastolic volume, RV ejection fraction, as survival predictors. We also trained a model on strain-related measures of mechanical function with tensors in the longitudinal, radial and circumferential directions[72]. A last model was trained on both the CMR parameters and a set of clinical risk factors[73], which comprised age, sex, six-minute walk distance, functional class and mean pulmonary artery pressure using the missForest algorithm to impute any missing values[74]. To account for collinearity among these predictor variables, a regularization term was added to the Cox partial likelihood function:

$$\log L(\beta) = \sum_{i=1}^{n}\delta_i\left\{\beta'\mathbf{x}_i - \log\sum_{j\in R(t_i)} e^{\beta'\mathbf{x}_j}\right\} + \frac{1}{2}\lambda\ \|\beta\|^2 \quad (10)$$

In the equation above, λ is a parameter that controls the strength of the penalty. λ was treated as a hyperparameter and its optimal value was selected via cross-validation. Internal validation of these models was carried out using the bootstrap-based procedure outlined in the previous section. Model comparisons were carried out using the R package survcomp[75] to compare concordance index measures (see Supplementary Information for further details).

Model interpretation. To facilitate interpretation of our 4Dsurvival network, we used Laplacian eigenmaps to project the learned latent code into two dimensions[8], allowing latent space visualization. Neural networks derive predictions through multiple layers of non-linear transformations on the input data. This complex architecture does not lend itself to straightforward assessment of the relative importance of individual input features. To tackle this problem, we used a simple regression-based inferential mechanism to evaluate the contribution of motion in various regions of the RV to the model's predicted risk. For each of the 202 vertices in our RV mesh models, we computed a single summary measure of motion by averaging the displacement magnitudes across 19 frames. This yielded one mean displacement value per vertex. This process was repeated across all subjects. Then we regressed the predicted risk scores onto these vertex-wise mean displacement magnitude measures using a mass univariate approach; that is, for each vertex v ($v=1, \ldots, 202$), we fitted a linear regression model where the dependent variable was predicted risk score, and the independent variable was average displacement magnitude of vertex v. Each of these 202 univariate regression models was fitted on all subjects and yielded 1 regression coefficient representing the effect of motion at a vertex on predicted risk. The absolute values of these coefficients, across all vertices, were then mapped onto a template RV mesh to provide a visualization of the differential contribution of various anatomical regions to predicted risk.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data and code availability

Algorithms, motion models and statistical analysis are publicly available on Github under a GNU General Public License (https://github.com/UK-Digital-Heart-Project/4Dsurvival)[76]. A training simulation is available as a Docker image with an interactive Jupyter notebook hosted on Code Ocean (https://doi.org/10.24433/CO.8519672.v1)[77]. Personal data are not available due to privacy restrictions.

Received: 8 October 2018; Accepted: 9 January 2019;
Published online: 11 February 2019

References

1. Wang, L., Zhao, G., Cheng, L. & Pietikäinen, M. *Machine Learning for Vision-Based Motion Analysis: Theory and Techniques* (Springer, London, 2010).
2. Mei, T. & Zhang, C. Deep learning for intelligent video analysis. *Microsoft*; https://www.microsoft.com/en-us/research/publication/deep-learning-intelligent-video-analysis/ (2017).
3. Liang, F., Xie, W. & Yu, Y. Beating heart motion accurate prediction method based on interactive multiple model: an information fusion approach. *Biomed. Res. Int.* **2017**, 1279486 (2017).
4. Savarese, G. & Lund, L. H. Global public health burden of heart failure. *Card. Fail. Rev.* **3**, 7–11 (2017).
5. Galie, N. et al. 2015 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur. Heart J.* **37**, 67–119 (2016).
6. Puyol-Antón, E. et al. A multimodal spatiotemporal cardiac motion atlas from MR and ultrasound data. *Med. Image Anal.* **40**, 96–110 (2017).
7. Scatteia, A., Baritussio, A. & Bucciarelli-Ducci, C. Strain imaging using cardiac magnetic resonance. *Heart Fail. Rev.* **22**, 465–476 (2017).

8. Belkin, M. & Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* 14 (eds Dietterich, T. G. et al.) 585–591 (MIT Press, Cambridge, 2002).
9. Li, K., Javer, A., Keaveny, E. E. & Brown, A. E. X. Recurrent neural networks with interpretable cells predict and classify worm behaviour. Preprint at https://doi.org/10.1101/222208 (2017).
10. Walker, J., Doersch, C., Gupta, A. & Hebert, M. An uncertain future: forecasting from static images using variational autoencoders. Preprint at https://arxiv.org/abs/1606.07873 (2016).
11. Bütepage, J., Black, M., Kragic, D. & Kjellström, H. Deep representation learning for human motion prediction and classification. Preprint at https://arxiv.org/abs/1702.07486 (2017).
12. Johnson, K. W. et al. Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl. Sci.* **2**, 311–327 (2017).
13. Cikes, M. & Solomon, S. D. Beyond ejection fraction: an integrative approach for assessment of cardiac structure and function in heart failure. *Eur. Heart J.* **37**, 1642–1650 (2016).
14. Ahmad, T. et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J. Am. Coll. Cardiol.* **64**, 1765–1774 (2014).
15. Shah, S. J. et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* **131**, 269–279 (2015).
16. Awan, S. E., Sohel, F., Sanfilippo, F. M., Bennamoun, M. & Dwivedi, G. Machine learning in heart failure: ready for prime time. *Curr. Opin. Cardiol.* **33**, 190–195 (2018).
17. Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K. & Fotiadis, D. I. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Comput. Struct. Biotechnol. J.* **15**, 26–47 (2017).
18. Ambale-Venkatesh, B. et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ. Res.* **121**, 1092–1101 (2017).
19. Yousefi, S. et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **7**, 11707 (2017).
20. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput. Biol.* **14**, 1–18 (2018).
21. Katzman, J. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
22. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
23. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine.*J. R. Soc. Interface* **15**, 20170387 (2018).
24. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
25. Shen, D., Wu, G. & Suk, H. I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
26. Bai, W. et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J. Cardiovasc. Magn. Reson.* **20**, 65 (2018).
27. Piras, P. et al. Morphologically normalized left ventricular motion indicators from MRI feature tracking characterize myocardial infarction. *Sci. Rep.* **7**, 12259 (2017).
28. Zhang, X. et al. Orthogonal decomposition of left ventricular remodeling in myocardial infarction. *Gigascience* **6**, 1–15 (2017).
29. Zhang, X. et al. Atlas-based quantification of cardiac remodeling due to myocardial infarction. *PLoS ONE* **9**, e110243 (2014).
30. Dawes, T. et al. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. *Radiology* **283**, 381–390 (2017).
31. Rifai, S., Vincent, P., Muller, X., Glorot, X. & Bengio, Y. Contractive auto-encoders: explicit invariance during feature extraction. In *Proc. 28th International Conference on Machine Learning*, 833–840 (Omnipress, 2011).
32. Rolfe, J. T. & LeCun, Y. Discriminative recurrent sparse auto-encoders. Preprint at 1301.3775 (2013).
33. Huang, R., Liu, C., Li, G. & Zhou, J. Adaptive deep supervised autoencoder based image reconstruction for face recognition. *Math. Probl. Eng.* **2016**, 14 (2016).
34. Du, F., Zhang, J., Ji, N., Hu, J. & Zhang, C. Discriminative representation learning with supervised auto-encoder. *Neur. Proc. Lett.* https://doi.org/10.1007/s11063-018-9828-2 (2018).
35. Zaghbani, S., Boujneh, N. & Bouhlel, M. S. Age estimation using deep learning. *Comp. Elec. Eng.* **68**, 337–347 (2018).
36. Beaulieu-Jones, B. K. & Greene, C. S. Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* **64**, 168–178 (2016).
37. Shakeri, M., Lombaert, H., Tripathi, S. & Kadoury, S. Deep spectral-based shape features for Alzheimer's disease classification. In *International Workshop on Spectral and Shape Analysis in Medical Imaging* (eds Reuter, M. et al.) 15–24 (Springer, 2016).
38. Biffi, C. et al. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* Vol. 11071 (eds Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C. & Fichtinger, G.) (Springer, 2018).
39. Dawes, T. J. W., Bello, G. A. & O'Regan, D. P. Multicentre study of machine learning to predict survival in pulmonary hypertension. *Open Science Framework* https://doi.org/10.17605/OSF.IO/BG6T9 (2018).
40. Grapsa, J. et al. Echocardiographic and hemodynamic predictors of survival in precapillary pulmonary hypertension: seven-year follow-up. *Circ. Cardiovasc. Imaging* **8**, 45–54 (2015).
41. Bao, W., Yue, J. & Rao, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE* **12**, e0180944 (2017).
42. Lim, B. & van der Schaar, M. Disease-atlas: navigating disease trajectories with deep learning. Preprint at https://arxiv.org/abs/1803.10254 (2018).
43. Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. DeepHit: a deep learning approach to survival analysis with competing risks. In *32nd Association for the Advancement of Artificial Intelligence (AAAI) Conference* (2018).
44. Gopalan, D., Delcroix, M. & Held, M. Diagnosis of chronic thromboembolic pulmonary hypertension. *Eur. Respir. Rev.* **26**, 160108 (2017).
45. Kramer, C., Barkhausen, J., Flamm, S., Kim, R. & Nagel, E. Society for cardiovascular magnetic resonance board of trustees task force on standardized protocols. Standardized cardiovascular magnetic resonance (CMR) protocols 2013 update. *J. Cardiovasc. Magn. Reson.* **15**, 91 (2013).
46. Woodbridge, M., Fagiolo, G. & O'Regan, D. P. MRIdb: medical image management for biobank research. *J. Digit. Imaging* **26**, 886–890 (2013).
47. Schulz-Menger, J. et al. Standardized image interpretation and post processing in cardiovascular magnetic resonance: society for cardiovascular magnetic resonance (SCMR) board of trustees task force on standardized post processing. *J. Cardiovasc. Magn. Reson.* **15**, 35 (2013).
48. Baggen, V. J. et al. Cardiac magnetic resonance findings predicting mortality in patients with pulmonary arterial hypertension: a systematic review and meta-analysis. *Eur. Radiol.* **26**, 3771–3780 (2016).
49. Hulshof, H. G. et al. Prognostic value of right ventricular longitudinal strain in patients with pulmonary hypertension: a systematic review and meta-analysis. *Eur. Heart J. Cardiovasc. Imaging* https://doi.org/10.1093/ehjci/jey120 (2018).
50. Duan, J. et al. Automatic 3D bi-ventricular segmentation of cardiac images by a shape-constrained multi-task deep learning approach. Preprint at 1808.08578 (2018).
51. Bai, W. et al. A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion. *Med. Image Anal.* **26**, 133–145 (2015).
52. Shi, W. et al. Temporal sparse free-form deformations. *Med. Image Anal.* **17**, 779–789 (2013).
53. Rueckert, D. et al. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* **18**, 712–721 (1999).
54. Bai, W et al. Learning a global descriptor of cardiac motion from a large cohort of 1000+ normal subjects. In *8th International Conference on Functional Imaging and Modeling of the Heart (FIMH'15)* Vol. 9126 (Springer, Cham, 2015).
55. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
56. Cox, D. Regression models and life-tables. *J. R. Stat. Soc. B* **34**, 187–220 (1972).
57. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
58. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge MA, 2016).
59. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995).
60. Abadi, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (TensorFlow, 2015); http://download.tensorflow.org/paper/whitepaper2015.pdf
61. Chollet, F. et al. *Keras* https://keras.io (2015).
62. Kennedy, J. & Eberhart, R. Particle swarm optimization. *Proc. IEEE Int. Conf. Neural Net.* **4**, 1942–1948 (1995).
63. Engelbrecht, A. *Fundamentals of Computational Swarm Intelligence* (Wiley, Chichester, 2005).
64. Lorenzo, P. R., Nalepa, J., Kawulok, M., Ramos, L. S. & Pastor, J. R. Particle swarm optimization for hyper-parameter selection in deep neural networks. In *Proc. Genetic and Evolutionary Computation Conference, GECCO '17*, 481–488 (2017).
65. Claesen, M., Simm, J., Popovic, D. & De Moor, B. Hyperparameter tuning in Python using Optunity.*In Proc. International Workshop on Technical Computing for Machine Learning and Mathematical Engineering* Vol. 9 (2014).

66. Harrell, F., Califf, R., Pryor, D., Lee, K. & Rosati, R. Evaluating the yield of medical tests.*J. Am. Med. Assoc.* **247**, 2543–2546 (1982).
67. Moons, K. et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–W73 (2015).
68. Harrell, F., Lee, K. & Mark, D. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
69. Efron, B. Estimating the error rate of a prediction rule: some improvements on cross-validation. *J. Am. Stat. Assoc.* **78**, 316–331 (1983).
70. Efron, B. & Tibshirani, R. in *An Introduction to the Bootstrap* Ch. 17 (Chapman & Hall, New York, 1993).
71. Smith, G., Seaman, S., Wood, A., Royston, P. & White, I. Correcting for optimistic prediction in small data sets. *Am. J. Epidem.* **180**, 318–324 (2014).
72. Liu, B. et al. Normal values for myocardial deformation within the right heart measured by feature-tracking cardiovascular magnetic resonance imaging. *Int. J. Cardiol.* **252**, 220–223 (2018).
73. Gall, H. et al. The Giessen pulmonary hypertension registry: survival in pulmonary hypertension subgroups. *J. Heart Lung. Transplant.* **36**, 957–967 (2017).
74. Stekhoven, D. J. & Buhlmann, P. missForest–non–parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2011).
75. Schroder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27**, 3206–3208 (2011).
76. Bello, G. A. & O'Regan, D. Deep learning cardiac motion analysis for human survival prediction (4Dsurvival) *Zenodo* https://doi.org/10.5281/zenodo.1451540 (2019).
77. Bello, G. et al. Deep learning cardiac motion analysis for human survival prediction (4Dsurvival). *Code Ocean* https://doi.org/10.24433/CO.8519672.v1 (2018).

Acknowledgements

The research was supported by the British Heart Foundation (NH/17/1/32725, RE/13/4/30184); the National Institute for Health Research Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London; and the Medical Research Council, UK. The TITAN Xp GPU used for this research was kindly donated by the NVIDIA Corporation.

Author contributions

G.A.B., C.B. and T.J.W.D. contributed to methodology, software, formal analysis and writing original draft. J.D. contributed to methodology, software and writing original draft; A.d.M. was involved with formal analysis; L.S.G.E.H., J.S.R.G., M.R.W. and S.A.C. were involved in investigation; D.R. contributed to software and supervision; D.P.O. was responsible for conceptualization, supervision, writing (review and editing) and funding acquisition. All authors reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0019-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.P.O.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

nature research

Corresponding author(s): D.P. O'Regan

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	Manual volumetric analysis of images (acquired from cardiac magnetic resonance imaging) was performed using the proprietary software cmr42 (Circle Cardiovascular Imaging, Calgary, Canada) Training and validation of deep learning and conventional parameter models was carried out using custom algorithms (available at: https://github.com/UK-Digital-Heart-Project/4Dsurvival) implemented with open-source Python language libraries Keras (v2.1.3 [http://keras.io]), Tensorflow-GPU (v1.4.0 [https://www.tensorflow.org/]), Optunity (v1.1.1 [https://github.com/claesenm/optunity]) and Lifelines (v0.14.6 [https://lifelines.readthedocs.io/en/latest/]).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Algorithms, motion models and statistical analysis are publicly available under a GNU General Public License. A training simulation is available as a Docker image with an interactive Jupyter notebook hosted on Binder. Personal data are not available due to privacy restrictions.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our power calculations use classifier performance estimates obtained from preliminary data in 256 PH patients comparing the sensitivity for identifying high-risk patients using supervised learning of cardiac motion versus conventional risk factors [Dawes et al. (2017) Radiology; 283(2):381-390]. A bootstrap cross-validation of our feasibility data using nested multivariable models demonstrated an incremental benefit of ML using complex phenotypes in outcome prediction (ANOVA, Hazard Ratio: F=80.2, p<0.001; AUC: F=94.2, p<0.001).
Data exclusions	Criteria for inclusion were a documented diagnosis of Group 4 pulmonary hypertension (PH) investigated by right heart catheterization (RHC) and non-invasive imaging. Subjects with congenital heart disease were excluded.
Replication	Results have not been replicated in an external cohort. The current work represents the preliminary stage of a multicentre study that will involve external validation in a future study (for further details, see published prospective study design: Dawes TJW, Bello GA, O'Regan DP. Multicentre study of machine learning to predict survival in pulmonary hypertension. Open Science Framework (2018). DOI 10.17605/OSF.IO/BG6T9 [https://osf.io/qvx69/])
Randomization	No experimental groups were used in this study
Blinding	Blinding is not relevant to this study, as we did not utilize experimental groups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Age (years): Mean=62.9, SD=14.5 Body surface area (m2): Mean=1.92, SD=0.25 Male: n=169 (56%); Female: n=133 (34%) Race: Caucasian 71.2%, Asian 2.3%, Black 4.3%, Other 9.3%, Unknown 12.9%

WHO functional class: Class I 0%, Class II 15%, Class III 71%, Class IV 14%
Systolic BP (mmHg): Mean=131.5, SD=25.2
Diastolic BP (mmHg): Mean=75, SD=13
Heart rate (beats/min): Mean=69.8, SD=22.5
Mean right atrial pressure (mmHg): Mean=9.9, SD=5.8
Mean pulmonary artery pressure (mmHg): Mean=44.1, SD=12.6
Pulmonary vascular resistance (Wood units): Mean=8.9, SD=5.0
Cardiac output (l/min): Mean=4.3, SD=1.5
LV ejection fraction (%): Mean=61, SD=11.1
LV end diastolic volume (ml): Mean=110, SD=37.4
LV end systolic volume (ml): Mean=44, SD=22.9
RV ejection fraction (%): Mean=38, SD=13.7
RV end diastolic volume (ml): Mean=194, SD=62
RV end systolic volume (ml): Mean=125, SD=59.3

Recruitment

This study was part of a continuous prospective research program into the prognosis of patients with PH by using conventional clinical and imaging biomarkers. Our study used data (cross-sectional) collected from patients referred to the National Pulmonary Hypertension Service (at the Imperial College Healthcare NHS Trust) for routine diagnostic assessment and cardiac imaging.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Structural imaging (Cardiac)

Field strength

1.5

Sequence & imaging parameters

A standard clinical protocol for cardiac MRI was followed according to published international guidelines (Kramer, JCMR 2013;15:91). Cardiac ventricular function was assessed using balanced-steady state free precession (b-SSFP) cine imaging acquired in conventional cardiac short- and long- axis planes typically with: repetition time msec/echo time msec, 3.2/1.6; voxel size, 1.5 x 1.5 x 8 mm; flip angle, 60°; sensitivity encoding factor (SENSE), 2.0; bandwidth, 962 Hz/pixel; temporal resolution 29 msec; slice thickness 10mm; field of view 400 x 400 mm, 30 time phases.

Area of acquisition

Protocolised, three-plane, low-resolution localizer images were used to define the ventricles as the region of interest, after which long- and short-axes planes were described using a line from the apex of the heart to centre of the left ventricular base. Margins of ~1cm in all planes were added to allow for variable breath-holding.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

Feature extraction was carried out via an image processing pipeline consisting of segmentation, co-registration and mesh generation. The output of this pipeline was a set of high-resolution, three-dimensional surface mesh representations of the heart's right ventricle (RV) at various phases of the cardiac cycle (total of 20 phases). These were used to derive point-wise displacement values representing the distance traveled by each mesh vertex (corresponding to an anatomical location on the RV) from frame to frame. These displacement values were fed as independent variables into a predictive neural network model. The neural network architecture used in this study was a 'supervised autoencoder', which combines dimension reduction with survival prediction via a Cox Proportional Hazards model. Training and validation metrics included the hazard ratio and Harrell's concordance index.