Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients

Gerhard-Paul Diller[1,2,3,4]*, Aleksander Kempny[1,2], Sonya V. Babu-Narayan[1,2], Marthe Henrichs[3], Margarita Brida[1,5], Anselm Uebing[1,6], Astrid E. Lammers[6], Helmut Baumgartner[3,4], Wei Li[1,2], Stephen J. Wort[1,2], Konstantinos Dimopoulos[1,2], and Michael A. Gatzoulis[1,2]

[1]Adult Congenital Heart Centre, National Centre for Pulmonary Hypertension, Royal Brompton Hospital, Sydney Street, SW3 6NP London, UK; [2]National Heart and Lung Institute, Imperial College School of Medicine, Dovehouse Street, SW3 6LY London, UK; [3]Department of Cardiology III – Adult Congenital and Valvular Heart Disease, University Hospital Muenster, Albert-Schweitzer-Campus 1, 48149 Muenster, Germany; [4]Competence Network for Congenital Heart Defects, DZHK (German Centre for Cardiovascular Research), Augustenburger Platz 1, 13353 Berlin, Germany; [5]Division of Valvular Heart Disease and Adult Congenital Heart Disease, Department of Cardiovascular Medicine, University Hospital Centre Zagreb, Kispaticeva 12, Zagreb, Croatia; and [6]Division of Paediatric Cardiology, University Hospital Muenster, Albert-Schweitzer-Campus 1, 48149 Muenster, Germany

See page 1078 for the editorial comment on this article (doi: 10.1093/eurheartj/ehz089)

Aims	To assess the utility of machine learning algorithms on estimating prognosis and guiding therapy in a large cohort of patients with adult congenital heart disease (ACHD) or pulmonary hypertension at a single, tertiary centre.
Methods and results	We included 10 019 adult patients (age 36.3 ± 17.3 years) under follow-up at our institution between 2000 and 2018. Clinical and demographic data, ECG parameters, cardiopulmonary exercise testing, and selected laboratory markers where collected and included in deep learning (DL) algorithms. Specific DL-models were built based on raw data to categorize diagnostic group, disease complexity, and New York Heart Association (NYHA) class. In addition, models were developed to estimate need for discussion at multidisciplinary team (MDT) meetings and to gauge prognosis of individual patients. Overall, the DL-algorithms—based on over 44 000 medical records—categorized diagnosis, disease complexity, and NYHA class with an accuracy of 91.1%, 97.0%, and 90.6%, respectively in the test sample. Similarly, patient presentation at MDT-meetings was predicted with a test sample accuracy of 90.2%. During a median follow-up time of 8 years, 785 patients died. The automatically derived disease severity-score derived from clinical information was related to survival on Cox analysis independently of demographic, exercise, laboratory, and ECG parameters.
Conclusion	We present herewith the utility of machine learning algorithms trained on large datasets to estimate prognosis and potentially to guide therapy in ACHD. Due to the largely automated process involved, these DL-algorithms can easily be scaled to multi-institutional datasets to further improve accuracy and ultimately serve as online based decision-making tools.
Keywords	Adult congenital heart disease • Prognostication • Mortality • Machine learning • Deep learning

* Corresponding author. Tel: +44 207 351 8602, Fax: +44 207 351 8629, Email: gerhard.diller@gmail.com

Introduction

Population-based studies suggest that adult patients with congenital heart disease (ACHD) under follow-up at tertiary centres and those managed according to current guidelines[1] have superior survival prospects compared with those not attending such institutions.[2,3] Therapy for ACHD is often empiric and based on retrospective data. With increasing numbers of ACHD patients and concentration of care at major tertiary centres combining data collected for routine care into risk stratification models is appealing. However, with growing volume of data, manual data collection and curation becomes logistically challenging as this approach is time consuming and expensive. Moreover, this is often unstructured information such as medical letters written in natural language that is unsuitable for direct inclusion in conventional statistical models. Machine learning solutions have shown promising results in other medical fields and might be applicable to ACHD patients.[4–6] The aim of the current study was to test the utility of deep neuronal networks in building prognostic models in synergy with an established statistical framework in a large ACHD cohort. Specifically, we aimed to test the hypothesis, that text analysis could allow for patient categorization into diagnostic subgroups, disease complexity subsets, and functional classes with comparable accuracy to manual data mining, representing the current gold standard method. In addition, we aimed to test the hypothesis that neuronal networks may allow to integrate medical reports directly into a prognostic model that is independent of manual data cleaning and can thus be extended to heterogenous multi-institutional datasets.

Patients and methods

We retrospectively collected data on all adult patients (age ≥16 years) under active follow-up at the Royal Brompton Hospital, London between 2000 and 2018. All clinic visits were identified from electronic health records and patient data including diagnosis, clinical status/symptoms, and medication were automatically extracted from medical letters using a dedicated text mining algorithm (for details Supplementary material online, *Appendix*). In addition, laboratory parameters, data from cardiopulmonary exercise testing and automatically collected ECG data were available. All patients were manually assigned to specific diagnostic subgroups and disease complexity (based on the Bethesda classification).[7] In addition, New York Heart Association (NYHA) class was manually coded for a random subset of 10 000 clinical letters to serve as input for supervised machine learning. Medication names derived from the national formulary were used to screen all reports for drug therapy employed. Laboratory, ECG, and cardiopulmonary exercise data were linked to clinical visits within 6 months of the respective tests. For ECG, exercise, and laboratory parameters imputation using the variable median was used to account for missing data. In addition, the records were screened for patient presentation at the joint medical/surgical multidisciplinary team (MDT) meeting. These data were subsequently used for supervised machine learning. Data on overall mortality were retrieved from the Office for National Statistics, which registers all United Kingdom deaths. As this was a retrospective analysis based on data collected for routine clinical care and administrative purposes (UK

National Research Ethics Service guidance), individual informed consent was not required.

Machine learning and statistical analysis

Descriptive data are presented as mean ± standard deviation or median and interquartile range (IQR) for continuous variables, whereas categorical variables are presented as number (%). The deep learning (DL)-models developed accept raw natural language data as input and categorize patients into binary (MDT and survival prediction) or categorical groups (diagnostic groups, disease complexity, and NYHA class). Specific network details are presented in the Supplementary material online, *Appendix*. The *Take home figure* and *Figure 1* show the model and the analysis workflow schematically. Briefly, the first layer of the DL-models consisted of an embedding layer accepting word tensors after tokenization and pre-processing. Subsequent layers included one-dimensional convolutional networks and long short-term memory layers. After concatenating the sub-models, the final layer consisted of a densely connected layer with sigmoid (for binary outcomes) or soft-max activation (for multiple categorical parameters), respectively. Hyperparameters were adjusted to ensure maximal accuracy while avoiding overfitting. Accuracy was calculated as previously described[6] on the test sample (20% of the original dataset), and loss was calculated using binary or categorical cross-entropy, respectively. Receiver operating characteristics (ROC) curve area for clustered data was used to account for multiple reports originating from one patient.[8] For the prognostic model, the class probabilities of the DL-model (*Take home figure*) trained on 5-year survival data was included in uni- and multivariate Cox-proportional hazard survival models after testing the proportional hazards assumption (by assessing the relationship between scaled Schoenfeld residuals and survival time), and analysis was limited to the test dataset (including 20% of the studied cohort). Multivariate models including age, ECG-derived, exercise, and laboratory data were built to minimize Akaike information criterion and metrics of discriminatory ability and calibration are presented. Networks were implemented based on Tensorflow (version 1.8) using the *keras* package (version 2.1.6) for R. Training and testing were performed on an Intel platform with GPU support (Nvidia GX 1070; Python version 3.5.5; CUDA version 9.0.176), and models and model weights were saved for further analysis. Analyses were performed using R-package version 3.5.1.[9] A two-sided *P*-value of <0.05 was considered indicative of statistical significance.

Results

Demographics and clinical background

We included 10 019 patients (49% females, mean age at baseline 36.3 ± 17.3 years, and NYHA class I/II/III–IV 87/8/5%, respectively) under active follow-up at our institution as illustrated in *Table 1*. Overall, 47% of patients had simple defects 26% moderate complexity and 16% complex congenital heart defects (the remaining 11% are not represented in the Bethesda classification). The results of the ECG analysis, laboratory measures used, and cardiopulmonary testing are presented in *Table 1*. The data covered an 18-year period including 63 326 medical letters. Of these, 44 421 reports had appropriate diagnostic fields, medication, and symptoms/clinical
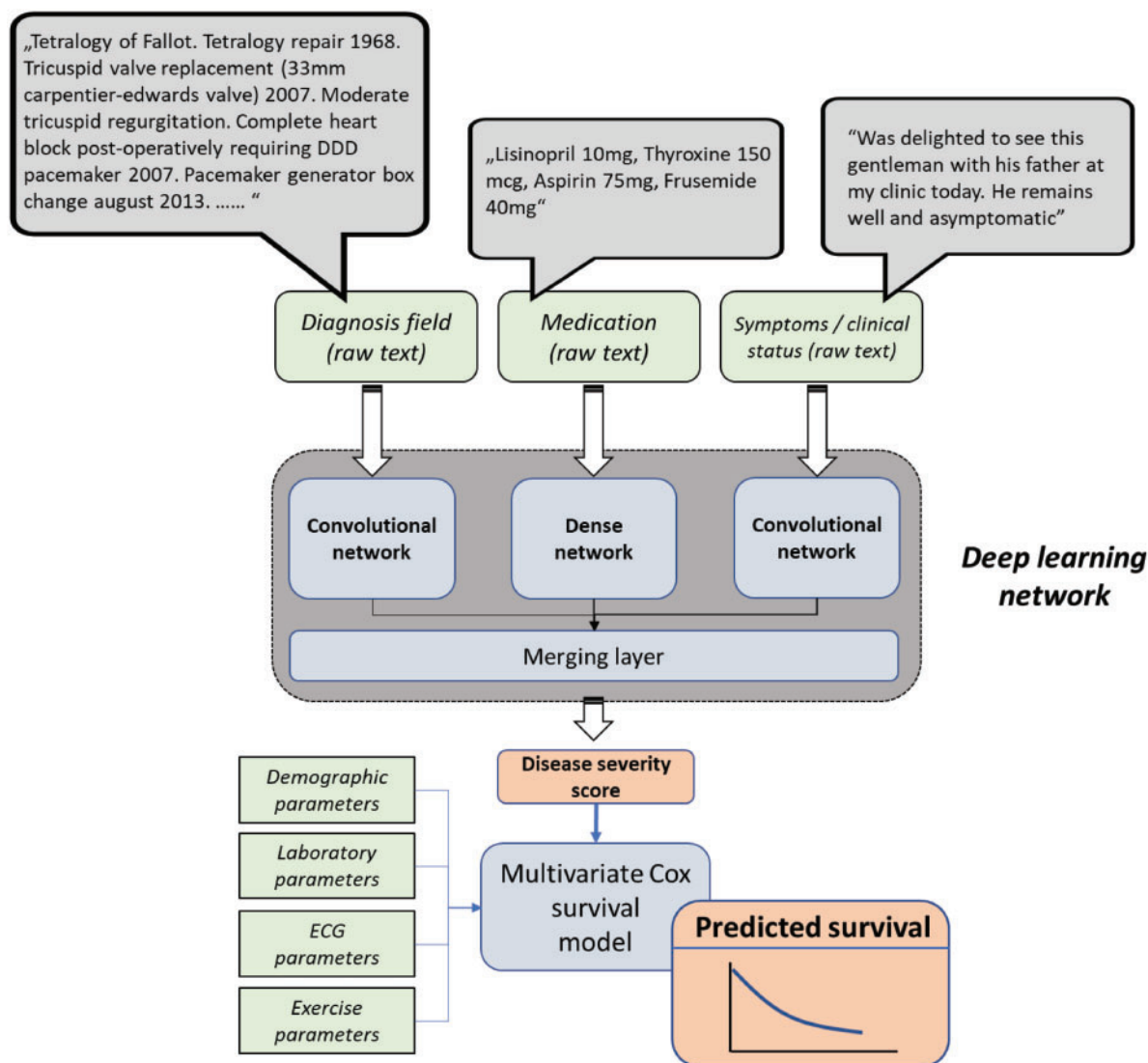
Take home figure Model overview illustrating the combination of deep learning architecture and semiparametric survival model. The deep learning network accepts raw text input and predicts main diagnostic group, disease complexity, and New York Heart Association class. In addition, a disease severity score modelled on 5-year mortality is provided that is combined with additional variables in a multivariate Cox model to provide prognostic information. The speech bubbles illustrate the type of input accepted by the deep learning network. Input examples shown are from different patients and slightly modified to avoid patient confidentiality issues.
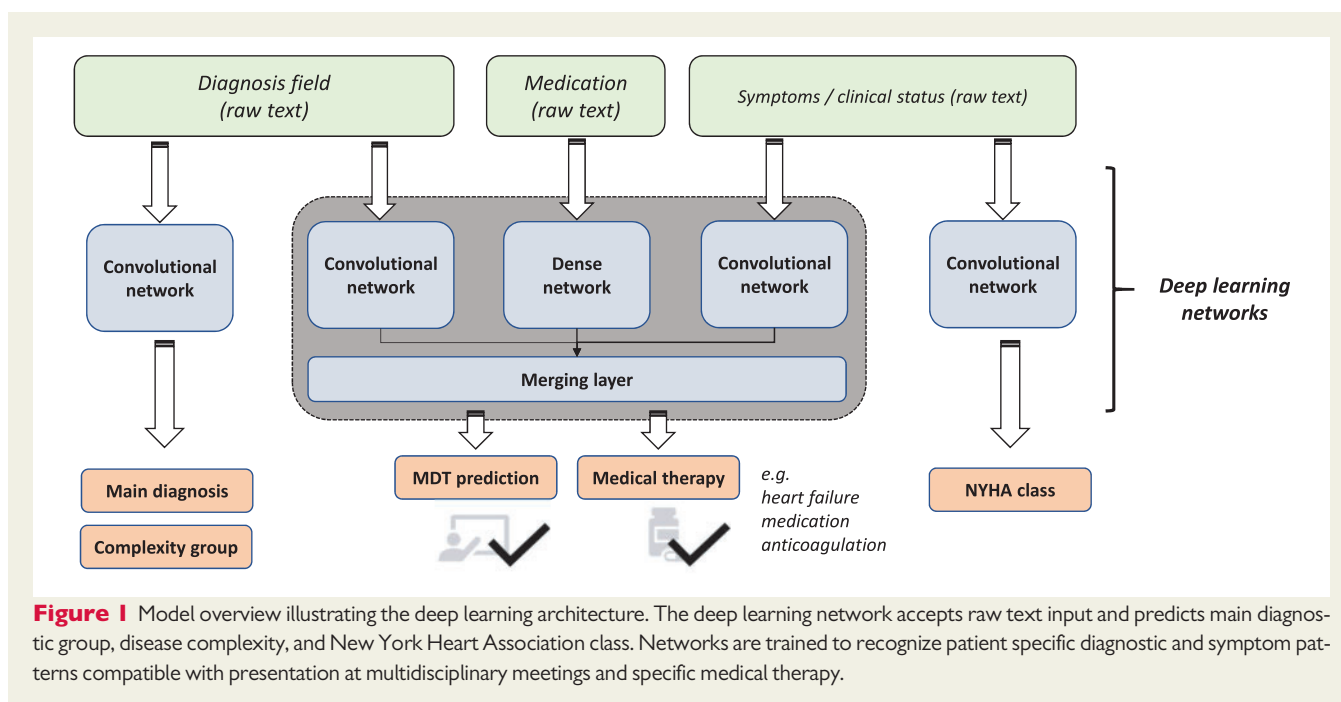
information and could be employed for analysis after initial cleaning. In addition, ECG data was available for 13 649 patient encounters, whereas laboratory and cardiopulmonary data was available for 14 307 and 10 502 time points, respectively.

Automatic classification of main diagnosis, disease complexity, and New York Heart Association class

Using raw text data derived from the diagnosis field of medical reports a DL-network was trained to predict main clinical diagnosis and disease complexity based on the Bethesda classification.

Punctuation, formatting, and spelling errors where deliberately not manually removed or corrected before training or testing. The diagnosis classification model achieved an accuracy of 91.1% in the test sample. This result was slightly lower when compared with manual classification by two experienced cardiologists based assessed on a subsample of 500 randomly selected cases (accuracy 96.3%). For disease complexity the model achieved a test accuracy of 97.0% (95% CI 96.7–97.3%). This was comparable to the accuracy score achieved when comparing manual coding between two cardiologists (G.-P.D. and A.E.L.; 98.2%).

The NYHA DL-model predicted the correct NYHA class with a test sample accuracy of 90.6% (95% CI 0.891–920). When presenting

Figure 1 Model overview illustrating the deep learning architecture. The deep learning network accepts raw text input and predicts main diagnostic group, disease complexity, and New York Heart Association class. Networks are trained to recognize patient specific diagnostic and symptom patterns compatible with presentation at multidisciplinary meetings and specific medical therapy.

a random collection of 500 text samples to two independent cardiologists (G.-P.D. and A.E.L.) an overall accuracy for determining NYHA class of 93.1% (95% CI 90.5–95.2%) was found.

Prediction of medical therapy and need for multidisciplinary team discussion

Over the study period, 2079 patients were presented at MDT meetings to discuss management, including surgical or catheter intervention or device implantation. A DL-model was built based on diagnosis, symptoms/clinical status, and medical treatment to predict the need for presentation at an MDT meeting within 6 months of the date of clinical presentation. The DL-model trained and tested with raw text data achieved an accuracy of 90.2% in the test sample (95% CI 89.3–91.0%) with an area under curve of 85.5% (95% CI 83.7–87.4%).

In addition, we constructed DL-models to predict treatment with various cardiac medication groups, including beta-blockers, ACE-inhibitors/ARBs, or anticoagulation based on diagnosis and symptoms/clinical presentation as well as other drugs administered. To avoid the model capturing information on the drug of interest in other parts of the medical report, any mention of the drug of interest or the medication group was automatically deleted from the input fields before analysis. *Table 2* illustrates the metrics obtained for predicting beta-blocker, ACE-inhibitor/angiotensin-receptor blocker, and anticoagulation in the train and test cohort. *Figure 2* shows the results of the ROC analysis for MDT classification and prediction of medication use.

Modelling all-cause mortality

During a median follow-up time of 8.05 years (IQR 6.34–13.26, corresponding to a total of 93 353 patient-years), 785 (7.6%) patients died. We constructed a DL-model including all available information,

linked to a downstream Cox-proportional hazard model to estimate mortality based on age, diagnosis, symptoms/clinical status, and medication (all presented to the DL-model as raw text), laboratory data, ECG parameters, and cardiopulmonary exercise data in the test dataset. The network architecture is presented in the *Take home figure* (for technical details see Supplementary material online, *Appendix*). As laboratory investigations, ECG and exercise parameters had a relevant proportion of missing values, data imputation with median values were applied to these parameters. On univariate Cox-analysis (using only the first available clinical visit per patient), the disease severity score of the DL-network was significantly related to all-cause mortality (hazard ratio for the disease severity score >0.9: 34.02; 95% CI 14.94–77.47, $P < 0.0001$) in the test sample. *Table 3* shows the results of the univariate Cox analysis for all significant parameters. Based on these results multivariate models were built in the test sample using backward elimination, based on minimizing Aikake criterion values. The final multivariate predictive model is shown in *Tables 4* and *5*.

Discussion

Our study shows that incorporating modern deep machine learning networks into prognostic models is feasible in congenital heart disease. By harvesting data from over 10 000 patients and 44 000 medical reports available at our tertiary centre for congenital heart disease and pulmonary hypertension and combining this information with demographic, laboratory, ECG, and exercise data, we were able to construct a powerful prognostic model that can be incorporated in electronic health record systems. The main advantage of our approach is that it obviates the need for human data cleaning and collection thus allowing to easily extend data sources in future projects. Furthermore, our model can be retrained and adjusted to new data

Table 1 Demographic and clinical characteristics by diagnostic group

Main diagnosis	Number	Male (%)	Age (years)	NYHA class (I/II/III) (%)	Peak heart rate (b.p.m.)	Peak VO2 (mL/kg/min)	VE/VCO2 slope	Peak systolic BP (b.p.m.)	Creatinine (µmol/L)	BNP (pg/mL)	Resting heart rate (b.p.m.)	QRS duration (ms)	QTc (ms)	Mortality (%)	DL-outcome score (z-score)
Aortic coarctation	991	59	33.9 ± 15.2	94/4/2	164 ± 25	29.2 ± 9.2	29.2 ± 6.8	175 ± 32	79.7 ± 22.1	17.7 ± 50.5	71.5 ± 14.9	105 ± 23	412 ± 39	3.4	−0.12
APVC	43	44	35.3 ± 19.1	93/6/1	164 ± 40	30.4 ± 10.4	32.8 ± 5.8	140 ± 23	71.1 ± 23.5	45.1 ± 25.2	72.2 ± 15.3	116 ± 32	457 ± 62	2.3	−0.14
ASD	1711	41	43.6 ± 17.4	83/13/4	153 ± 34	24.7 ± 9.4	33.4 ± 8.5	150 ± 26	78.8 ± 25.4	39.1 ± 73.7	75.7 ± 15.6	104 ± 27	426 ± 54	4.7	−0.08
AVSD	313	44	32.8 ± 15.7	91/4/5	155 ± 29	24.3 ± 8.2	31.5 ± 8.5	150 ± 21	82.1 ± 55.5	37.9 ± 40.4	81.7 ± 22.3	118 ± 31	443 ± 123	3.6	−0.08
Cardiomyopathy	154	56	37.2 ± 17.3	75/10/15	150 ± 31	23.4 ± 9.4	34.5 ± 8.5	139 ± 28	90.9 ± 34.6	73.4 ± 197.1	73.9 ± 22.6	123 ± 38	435 ± 59	19.5	0.28
CCTGA	128	52	37.7 ± 15.8	84/10/6	139 ± 32	20.4 ± 8.0	39.4 ± 13.6	143 ± 25	89.5 ± 35.2	66.7 ± 77.7	73.5 ± 14.0	126 ± 31	454 ± 60	17.7	0.19
Complex	437	51	30.4 ± 13.8	74/15/11	143 ± 33	20.1 ± 8.3	42.8 ± 17.5	142 ± 24	88.1 ± 46.1	95.4 ± 341.3	76.9 ± 17.0	126 ± 34	448 ± 75	19.5	0.37
Ebstein	181	45	39.6 ± 17.1	79/14/7	160 ± 27	22.2 ± 7.9	37.5 ± 12.8	150 ± 24	83.8 ± 23.7	14.5 ± 13.1	73.9 ± 15.1	126 ± 31	444 ± 36	7.3	−0.09
Eisenmenger	388	36	35.4 ± 13.7	25/40/35	133 ± 26	13.0 ± 4.7	67.8 ± 45.5	135 ± 20	85.1 ± 35.9	38.6 ± 58.5	76.8 ± 15.5	105 ± 21	437 ± 40	23.6	0.35
Fontan	162	48	27.1 ± 10.9	77/14/9	146 ± 33	21.2 ± 8.3	38.5 ± 11.7	136 ± 18	90.9 ± 42.8	29.8 ± 44.4	79.5 ± 70.5	104 ± 21	440 ± 101	15.9	0.23
PAH	213	47	56.5 ± 16.9	14/51/35	146 ± 28	15.3 ± 4.9	54.5 ± 12.6	152 ± 38	87.9 ± 38.6	72.9 ± 100.1	75.0 ± 15.2	107 ± 16	461 ± 38	35.2	0.28
Marfan	344	59	35.5 ± 16.5	98/2/0	157 ± 30	25.5 ± 9.9	31.7 ± 11.8	152 ± 31	71.8 ± 19.9	17.8 ± 17.1	68.8 ± 18.8	105 ± 23	417 ± 41	5.8	0.11
Other	406	52	37.9 ± 18.4	86/10/4	157 ± 31	26.0 ± 10.0	32.0 ± 9.4	147 ± 30	81.3 ± 30.3	41.2 ± 53.5	73.5 ± 14.2	103 ± 22	425 ± 33	4.4	−0.12
PDA	145	29	36.5 ± 18.0	86/7/7	169 ± 26	26.4 ± 10.2	30.7 ± 7.4	145 ± 24	67.4 ± 17.3	28.9 ± 62.6	70.6 ± 21.9	83 ± 23	381 ± 100	1.4	−0.19
Tetralogy of Fallot	1018	54	32.8 ± 15.2	90/7/3	163 ± 28	25.0 ± 8.5	32.6 ± 11.5	148 ± 24	81.2 ± 30.4	25.9 ± 68.6	73.9 ± 13.5	146 ± 29	460 ± 46	5.3	−0.1
TGA arterial switch	210	60	22.3 ± 6.6	95/5/0	177 ± 24	32.2 ± 9.8	30.4 ± 5.3	156 ± 25	72.9 ± 17.1	11.1 ± 10.1	75.2 ± 15.3	108 ± 19	427 ± 35	1.0	−0.21
TGA atrial switch	191	60	31.0 ± 12.4	84/10/6	153 ± 29	21.5 ± 7.1	38.0 ± 12.4	139 ± 20	74.0 ± 22.0	31.2 ± 33.3	69.2 ± 16.2	107 ± 29	439 ± 41	5.6	−0.13
TGA Rastelli	19	74	29.8 ± 9.3	93/7/0	174 ± 23	23.9 ± 6.4	34.3 ± 12.4	154 ± 26	92.8 ± 51.2	44.9 ± 34.3	77.9 ± 11.9	151 ± 14	463 ± 24	15.0	−0.11
Truncus arteriosus	19	47	25.9 ± 9.1	88/11/1	167 ± 28	25.3 ± 8.6	33.33 ± 8.8	148 ± 26	81.3 ± 28.8	18.8 ± 9.9	79.8 ± 12.0	126 ± 35	459 ± 39	15.8	0.19
Valvar	2055	55	37.6 ± 19.3	91/6/3	161 ± 29	28.1 ± 9.6	31.6 ± 9.4	154 ± 26	83.8 ± 46.4	30.9 ± 47.4	73.3 ± 14.9	103 ± 26	421 ± 45	7.9	−0.07
VSD	840	50	30.9 ± 14.1	93/4/3	160 ± 30	28.5 ± 9.9	30.2 ± 8.4	151 ± 28	79.2 ± 30.3	27.5 ± 49.4	73.6 ± 15.1	118 ± 32	434 ± 46	2.3	−0.17
VSD/PDA	51	55	28.6 ± 12.1	97/3/0	167 ± 13	21.8 ± 6.1	32.1 ± 4.4	146 ± 29	83.9 ± 14.1	30.2 ± 34.3	73.9 ± 11.7	113 ± 30	432 ± 34	2.0	−0.17
Total	10 019	51	36.3 ± 17.3	87/8/5	156 ± 31	25.2 ± 9.6	35.4 ± 16.2	149 ± 29	82.9 ± 36.4	40.9 ± 103.0	74.5 ± 27.6	115 ± 31	435 ± 58	7.6	0.00

APVC, anomalous pulmonary venous connection; ASD, atrial septal defect; AVSD, atrioventricular septal defect; BNP, brain natriuretic peptide; ccTGA, congenitally corrected TGA; PAH, pulmonary arterial hypertension; PDA, persistent arterial duct; TGA, transposition of the great arteries; VO2, oxygen uptake; VSD, ventricular septal defect.

Table 2 Accuracy and area under curve on receiver operating curve analysis for deep learning models predicting patient medication based on diagnosis, symptoms/clinical status, and other drugs utilized

Drug therapy	Accuracy	(95% CI)	AUC	(95% CI)
Beta-blocker				
Training cohort	0.943	0.940–0.945	0.966	0.963–0.969
Test cohort	0.892	0.886–0.899	0.859	0.845–0.873
ACE inhibitor/ARB				
Training cohort	0.955	0.952–0.957	0.973	0.970–0.976
Test cohort	0.891	0.885–0.898	0.897	0.886–0.907
Anticoagulation				
Training cohort	0.9606	0.959–0.963	0.980	0.977–0.982
Test cohort	0.908	0.902–0.914	0.912	0.902–0.923

sources without major effort and, thus, could eventually be employed to predict outcome and provide clinical guidance based on national or international ACHD datasets.

Unlike traditional statistical methods neuronal network-based systems can accept diverse data sources as input variables.[4,10] This considerably broadens the databases available and the scope for building prognostic models. For example, it is conceivable that data on exercise and pulse rate information collected by wearable devices could—at least in part—replace formal exercise tests in future. In addition, collecting and aggregating data on patient symptoms could reduce the number of clinic visits at tertiary centres and, thus make clinic visits more efficient by reducing time spend by physicians and other health care professionals in collecting such information. Broberg *et al.*[11] have highlighted the need and challenge in automated data acquisition from electronic health care records in ACHD patients in 2015. Over the past 3 years the area of artificial intelligence has seen major advances in image recognition, text classification, and generative models. This now allows us to utilize personal computers equipped with graphic processing units to perform these tasks and overcome most of the previously highlighted challenges of unstructured data by Broberg. By harvesting this data it should be possible to tracking quality of care in ACHD, which is likely to improve delivery of care.[11,12]

Artificial intelligence and machine learning models are increasingly used in cardiology. These data rich technologies should allow to augment and extend the effectiveness of cardiologists in future.[4] It has been argued that these tools will enable quicker, more efficient, and personalized care by incorporating various data sources, including genomic information, streamed mobile device data as well as data from outside the health care sector such as social media. Although certainly not a panacea for prediction model development, machine learning tools offer the promise of being more generalizable and applicable to external datasets.[4] While initially mainly applied to image classification, ECG diagnosis, or segmentation tasks,[4–6,13–16] machine learning tools are increasingly being utilized for building clinical prediction models in cardiovascular disease.[5,10] This includes the effort to develop risk stratification models for heart failure and arrhythmic

events as well as pilot studies attempting to predict risk of congenital heart surgery with machine learning algorithms.[10,17–20] The current study illustrates how such models may be applied in the setting of a life-long chronic disease such as ACHD.

In future, multicentre efforts are desirable to pool data and allow for training and testing of large comprehensive DL-networks. This raises obvious issues with privacy regulations and data anonymization. Therefore, supra-regional registries or networks supporting local centres and responsible for database maintenance and managing all the identifying medical information are required. In addition, it is conceivable that software tools for data acquisition, anonymization, and encryption could be deployed locally subsequently allowing to combine multi-institutional records with outcome data.

Strength of the current report

To the best of our knowledge, this is the largest single cohort study assessing prognosis in ACHD patients. With a total of over 90 000 patient-years included it even compares favourably to most registry-based efforts. In addition, by using deep machine learning networks, we provide a flexible framework that can be extended to additional centres and registries overcoming previous barriers. It also provides interfaces for including additional data (e.g. imaging raw data, data from wearable devices etc.). By minimizing human labour effort, efficiency is increased, and subjective bias is reduced.

Limitations

As a single centre retrospective study, the patients forming the basis of the current analysis may not necessarily be representative of the pattern of ACHD present in the community. A structured letter format, including a list of diagnoses is required for the text mining algorithm utilized. This is, however, standard practice in tertiary centres nowadays and essential to communicate clinical information to general practitioners and other relevant health care professionals. Models recognizing treatment strategies and guiding cardiologists based on previous therapeutic decisions may not necessarily propose the most beneficial/optimal strategy. This type of analysis does, therefore, not obviate the need for prospective clinical trials. However, by emulating expert decisions guided by decade long clinical experience, we contend that network models trained on similar data could prevent catastrophic treatment mistakes by alerting health care professionals that the care provided for individual patients may not be in line with the pattern statistically expected. For logistic reasons, we limited the number of laboratory and ECG parameters included in the analysis to those demonstrated by previous studies to be related to outcome in ACHD. Further studies may extend on this, and also include novel neurohormonal makers, inflammatory parameters, or markers of autonomic function in the analysis.

Including raw imaging sequences (e.g. echocardiographic or cardiac magnetic resonance imaging data) directly into the model is technically feasible, requires, however, adequately trained and diagnosis specific networks. As these specifically trained networks are currently still under development further studies including comprehensive clinical and imaging data are required to test their prognostic value in ACHD patients. We cannot exclude the possibility that prediction of MDT or medication use may capitalize on specific wording and subtle verbal hints included in the medical reports. However, as the model
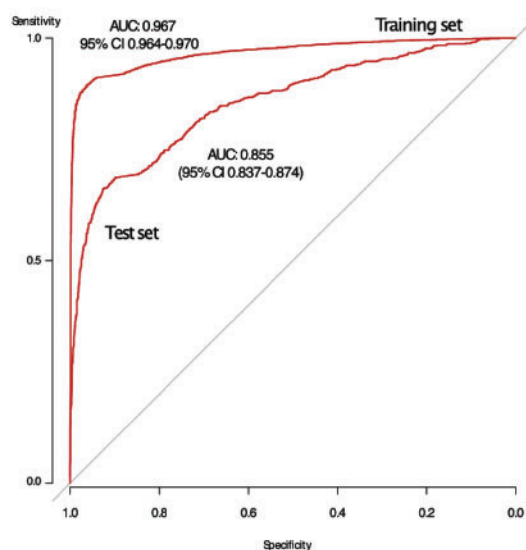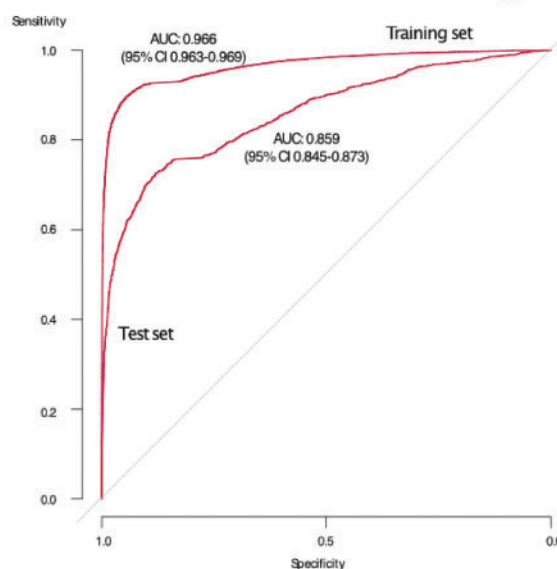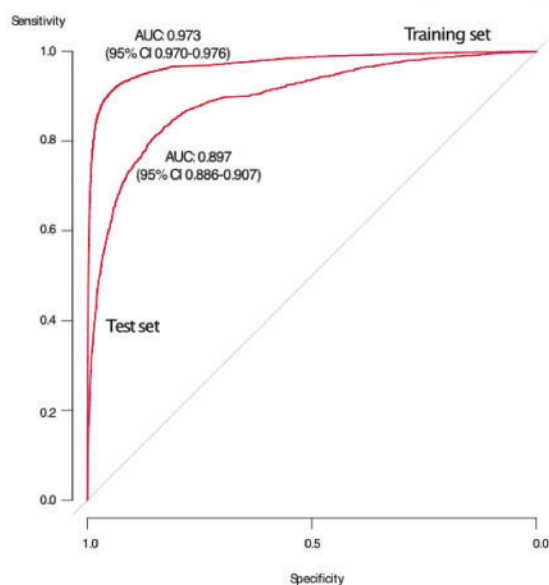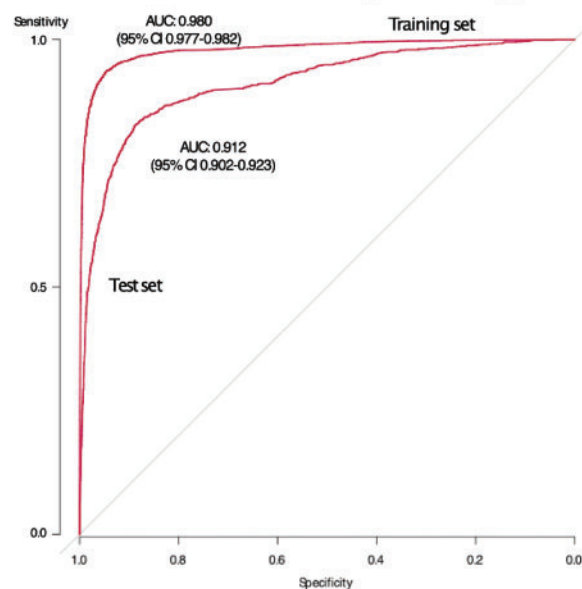
Figure 2 Results of the receiver operating curve analysis for predicting need for multidisciplinary discussion, beta-blocker, ACE-inhibitor/angiotensin receptor blocker, and anticoagulation in the training and validation sample, respectively.

included information on diagnosis, symptoms and medication, we contend that it is plausible that it identified especially highly symptomatic patients with high complexity lesions who may be more likely to be listed for MDT or treated with specific drugs. Ultimately, however, this illustrates the need for future testing of the model on external data.

Direct text analysis is language specific, therefore, the model must be adapted and retrained before use in a different language setting, limiting the generalizability of the model. Lastly, our model awaits external validation.

Conclusions

Our data illustrates the utility of machine learning algorithms trained on large datasets to estimate prognosis and potentially guide therapy in ACHD. Due to the largely automated process involved, these methods can easily be scaled to multi-institutional datasets to further improve accuracy and ultimately serve as online decision-making tools. Harvesting the enormous source of data machine learning can provide without affecting patients' privacy will represent a leading research challenge in the years to come. However, with appropriate

Table 3 Results of the univariate Cox proportional hazard analysis

| Parameter | Univariate model | | | |
	Hazard ratio	95% CI	P-value	Concordance
Deep learning disease severity score >0.9	34.018	14.940–77.470	<0.001	0.73
Age (10 years)	1.485	1.431–1.541	<0.001	0.69
Gender (male = 1)	1.335	1.156–1.540	<0.001	0.54
ECG parameters				
Resting heart rate (b.p.m.)	1.015	1.007–1.023	<0.001	0.58
QRS duration (ms)	1.010	1.007–1.014	<0.001	0.62
QTc duration (ms)	1.009	1.007–1.012	<0.001	0.63
Laboratory parameters				
Creatinine	1.007	1.006–1.008	<0.001	0.66
Brain natriuretic peptide	1.003	1.001–1.004	<0.001	0.76
Exercise parameters				
Peak heart rate (b.p.m.)	0.974	0.969–0.979	<0.001	0.77
Peak oxygen uptake (mL/kg/min)	0.883	0.858–0.909	<0.001	0.79
Peak systolic blood pressure (mmHg)	0.984	0.975–0.992	<0.001	0.64
VE/VCO$_2$ slope	1.035	1.026–1.044	<0.001	0.74

Hazard ratios (HR), 95% confidence interval, *P*-values, and concordance statistics are presented.

Table 4 Results of the multivariate Cox proportional hazard analysis: conventional clinical parameters only

| Parameter | Multivariate model (without disease severity score) | | | |
	Hazard ratio	95% CI	P-value	Concordance
Age (10 years)	1.258	1.108–1.428	<0.001	
Gender (male = 1)	1.658	1.055–2.607	0.03	
ECG parameters				
QRS duration (ms)	1.014	1.009–1.019	<0.001	
Laboratory parameters				
Creatinine	1.005	1.0005–1.0097	0.03	
Exercise parameters				
Peak heart rate (b.p.m.)	0.991	0.983–0.998	0.009	
Peak oxygen uptake (mL/kg/min)	0.918	0.885–0.951	<0.001	
Peak systolic blood pressure (mmHg)	0.985	0.976–0.994	0.001	
				0.82

Hazard ratios (HR), 95% confidence interval, *P*-values, and concordance statistics are presented.

Table 5 Results of the multivariate Cox proportional hazard analysis: including deep learning disease severity score (in the test sample only)

| Parameter | Multivariate model (including disease severity score) | | | |
	Hazard ratio	95% CI	P-value	Concordance
Disease severity score >0.9	34.02	14.94–77.47	<0.001	
Age (10 years)	1.065	1.047–1.084	<0.001	
Gender (male = 1)	2.738	1.412–5.309	0.003	
Laboratory parameters				
BNP	1.007	1.001–1.012	0.01	
				0.85

Hazard ratios (HR), 95% confidence interval, *P*-values, and concordance statistics are presented.

investment, a nationwide risk stratification and therapy guiding model could be provided to all health care professionals likely improving care for ACHD patients and helping to avoid potential catastrophic treatment errors in this vulnerable patient population.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

References

1. Baumgartner H, Bonhoeffer P, De Groot NM, de Haan F, Deanfield JE, Galie N, Gatzoulis MA, Gohlke-Baerwolf C, Kaemmerer H, Kilner P, Meijboom F, Mulder BJ, Oechslin E, Oliver JM, Serraf A, Szatmari A, Thaulow E, Vouhe PR, Walma E; Task Force on the Management of Grown-up Congenital Heart Disease of the European Society of Cardiology (ESC); Association for European Paediatric Cardiology (AEPC); ESC Committee for Practice Guidelines (CPG). ESC Guidelines for the management of grown-up congenital heart disease (new version 2010). *Eur Heart J* 2010;**31**:2915–2957.
2. Cordina R, Nasir Ahmad S, Kotchetkova I, Eveborn G, Pressley L, Ayer J, Chard R, Tanous D, Robinson P, Kilian J, Deanfield JE, Celermajer DS. Management errors in adults with congenital heart disease: prevalence, sources, and consequences. *Eur Heart J* 2018;**39**:982–989.
3. Mylotte D, Pilote L, Ionescu-Ittu R, Abrahamowicz M, Khairy P, Therrien J, Mackie AS, Marelli A. Specialized adult congenital heart disease care: the impact of policy on mortality. *Circulation* 2014;**129**:1804–1812.
4. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;**71**:2668–2679.
5. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017;**69**:2657–2664.
6. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med* 2018;**1**:6.
7. Williams RG, Pearson GD, Barst RJ, Child JS, del Nido P, Gersony WM, Kuehl KS, Landzberg MJ, Myerson M, Neish SR, Sahn DJ, Verstappen A, Warnes CA, Webb CL; National Heart Lung, and Blood Institute Working Group on research in adult congenital heart disease. Report of the National Heart, Lung, and Blood Institute Working Group on research in adult congenital heart disease. *J Am Coll Cardiol* 2006;**47**:701–707.
8. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997;**53**:567–578.
9. R Developement Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2010. http://www.R-project.org (17 November 2018).
10. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, van Rosendael AR, Beecy AN, Berman DS, Leipsic J, Nieman K, Andreini D, Pontone G, Schoepf UJ, Shaw LJ, Chang HJ, Narula J, Bax JJ, Guan Y, Min JK. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J* 2018.
11. Broberg C, Sklenar J, Burchill L, Daniels C, Marelli A, Gurvitz M. Feasibility of using electronic medical record data for tracking quality indicators in adults with congenital heart disease. *Congenit Heart Dis* 2015;**10**:E268–E277.
12. Gurvitz M, Marelli A, Mangione-Smith R, Jenkins K. Building quality indicators to improve care for adults with congenital heart disease. *J Am Coll Cardiol* 2013;**62**:2244–2253.
13. Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, Johnson AE, Mark RG. AF classification from a short single lead ECG recording: the PhysioNet/computing in cardiology challenge 2017. *Comput Cardiol (2010)* 2017;**44**. doi: 10.22489/CinC.2017.065-469. Epub 2018 April 5.
14. Luo G, An R, Wang K, Dong S, Zhang H. A deep learning network for right ventricle segmentation in short-axis MRI. *Comput Cardiol* 2016;**43**:485–488.
15. Pace DF, Dalca AV, Geva T, Powell AJ, Moghari MH, Golland P. Interactive whole-heart segmentation in congenital heart disease. *Med Image Comput Comput Assist Interv* 2015;**9351**:80–88.
16. Poh MZ, Poh YC, Chan PH, Wong CK, Pun L, Leung WW, Wong YF, Wong MM, Chu DW, Siu CW. Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms. *Heart* 2018;**104**:1921–1928.
17. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, Bluemke DA, Lima JAC. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circ Res* 2017;**121**:1092–1101.
18. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc* 2018;**7**. pii: e008678. doi: 10.1161/JAHA.118.008678.
19. Ruiz-Fernandez D, Monsalve Torra A, Soriano-Paya A, Marin-Alonso O, Triana Palencia E. Aid decision algorithms to estimate the risk in congenital heart surgery. *Comput Methods Programs Biomed* 2016;**126**:118–127.
20. Dudchenko A, Kopanitsa G. Decision support systems in cardiology: a systematic review. *Stud Health Technol Inform* 2017;**237**:209–214.