DEVELOPMENT AND VALIDATION OF **MACHINE LEARNING** MODELS IN CARDIOLOGY

RICARDO RICCI LOPES

DEVELOPMENT AND VALIDATION OF **MACHINE LEARNING** MODELS IN CARDIOLOGY

RICARDO RICCI LOPES

Development and Validation of Machine Learning Models in Cardiology

Ricardo Ricci Lopes

Layout: Ricardo Ricci Lopes
Cover Design: Jonas Marques
Print: Ridderprint | www.ridderprint.nl

ISBN: 978-94-6483-014-9

The work in this thesis was supported by ITEA3 Partner: Project 16017.

Financial support by the Dutch Heart Foundation for publication of this thesis is gratefully acknowledged.

Development and Validation of Machine Learning Models in Cardiology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 19 april 2023, te 10.00 uur

door Ricardo Ricci Lopes
geboren te Batatais/SP

Promotiecommissie

Promotores:	prof. dr. H.A. Marquering	AMC-UvA
	prof. mr. dr. B.A.J.M. de Mol	AMC-UvA
Overige leden:	prof. dr. I. Išgum	AMC-UvA
	prof. dr. M.C. Schut	Vrije Universiteit Amsterdam
	prof. dr. J. Kluin	AMC-UvA
	prof. dr. A. Abu-Hanna	AMC-UvA
	prof. dr. R.J. de Winter	AMC-UvA
	prof. dr. ir. H. Boersma	Erasmus Universiteit Rotterdam

Faculteit der Geneeskunde

Table of Contents

1

Introduction

Cardiovascular diseases

Cardiovascular diseases (CVD) are the leading cause of death globally, according to the World Health Organization. Most of the CVD can be prevented with a healthy lifestyle and controlled use of harmful substances, such as tobacco and alcohol (1). Early diagnosis and appropriate treatment are important factors to reduce mortality ratio of CVD. Prognostic and diagnostic models have been developed to support doctors and patients in the decision-making process on medical treatment or surgical interventions. For instance, multiple prognostic risk scores have been developed to estimate the risk of cardiac surgeries. Examples of these models are the Society Thoracic Surgeon (STS) risk model and the EuroScore I and II models, which generate a risk profile of adult surgical patients (2,3). These models have been generated several years ago; the STS was developed in 2008 and was updated in 2018, the EuroScore was developed in 2008 and was updated in 2012. It has been suggested that the methodology used to create these risk scores is outdated since it used traditional statistical approaches, however, methods improved accuracy could be achieved with machine learning (ML) models (2–5).

Considering that treatments and procedures are advancing over the years, the prognostic risk scores must also be regularly updated and evaluated over time. Moreover, these models do not present the same accuracy over different populations (5–8). In addition, new ML techniques have emerged as the state of the art for prediction modelling (9,10). These techniques, as well as the traditional ML techniques, are currently being deeply explored and evaluated in the cardiovascular field (2,11–13). Hence, this thesis supports the decision-making process in four cardiovascular diseases, which are presented as follows with a description of the relevant prognostic or diagnostic models and their main limitations.

Aortic valve stenosis

The human heart has four valves, which are responsible to keep blood moving in the right direction. Defects on these valves, developed or congenital, might cause

severe problems such as heart failure or death (14). Aortic stenosis (AS) is the most common valve disease requiring surgery in Europe and North America (15,16), with a prevalence of 1.7% in the general population aged over 65 years in developed countries (16). Over time, it is common that calcium from the blood deposits on the valve, causing calcifications in the aortic leaflets and malfunction of the valve, forcing the heart to work harder to compensate for it (17). AS is more common in the elderly population and it is usually associated with fatigue, dizziness, chest pain, and shortness of breath. Without surgery, up to 50% of the patients with severe AS die within one year (18). Among treatment options, transcatheter aortic valve implantation (TAVI) has emerged as a minimally invasive procedure and has been used as the last resort for high-risk patients considered inoperable (19). Currently, TAVI is also considered a viable treatment for low and intermediate-risk AS patients (20).

Multiple risk scores can be used to assess risks and support the decision-making process on TAVI (21). The STS and EuroScore II were not specifically created for TAVI, although are commonly assessed for TAVI procedures. For in-hospital and early mortality, the STS/ACC TAVR (22) and France2 (23) scores have been developed. However, in the general population, these scores do not maintain the same accuracy as in the original population for which they were created (24,25). This highlights the importance of developing treatment models specifically for a target population.

Coronary Artery Disease

Coronary arteries provide blood to the heart muscle. Over time, a build-up of plaques made up of cholesterol can cause narrowing of the coronary arteries. The excessive plaque build-up, called Coronary Artery Disease (CAD), reduces the blood flow and might cause heart attacks (26). CAD is the most common type of heart disease, being the leading cause of death in the United States (27) and being responsible for approximately 20% of deaths in Europe (28).

Computed tomographic coronary angiography (CTCA) is a non-invasive imaging technique used to support the diagnosis of CAD (29,30). For accurate assessment of the disease on CTCA, a minimal intra-arterial attenuation value is recommended (31). However, the contrast material adjustment for the CTCA

acquisition, in some cases, is not optimal. With that, the coronary attenuation values might be too low, thereby jeopardizing the diagnostic value of CTCA (32). Although there are several ML models to support the diagnosis of CAD (33), there are no studies aiming at the prediction of CTCAs with insufficient attenuations before its acquisition.

Atrial fibrillation

The heart's rhythm is controlled by an electrical signal which normally starts in the sinus node, travels through the atrioventricular node and bundle of His to the ventricles. A problem with this electrical signal may cause cardiac arrhythmia, making the heart to beat too fast, too slow, or irregularly (34). Atrial fibrillation (AF) is the most common cardiac arrhythmia in adults, with a prevalence of 2-4% (35). AF is strongly associated with a variety of conditions, such as valve diseases, coronary artery disease, cardiomyopathies, and stroke (36). The standard for the diagnosis of AF is by detection on an electrocardiogram (ECG). In patients with AF, the ECG is characterized by the absence of P-waves and irregular R intervals in the heart signal.

Cases of AF can be frequent and recurrent, requiring treatment with drugs or medical procedures (37). A well-established procedure for AF prevention is catheter ablation (37,38). This procedure consists of damaging specific regions of the heart that are causing arrhythmia to prevent new cases of AF. Although effective, with a rate of 69-80% of non-AF recurrence (39), the patient is not always AF-free after the procedure, which could even increase the risk of comorbidities and cause an impaired quality of life.

The use of ML techniques has improved automated detection of AF on ECGs (40,41). AF has a clear pattern on the ECG, and the most advanced ML techniques can learn important patterns automatically, reaching high accuracy in the detection of visual patterns. However, the recurrence of AF after the ablation procedure does not have a known pattern in the ECG, making its detection less accurate. Although there are some known risk factors and risk scores available (42–44), there are still no well-accepted ML models for AF recurrence detection. Investigations in this field can lead to discoveries and insights.

Cardiomyopathies

Cardiomyopathies are a group of structural heart diseases associated with the malfunctioning of the heart's muscles. There are many causes for cardiomyopathies, such as genetic factors, coronary disease (ischemic) and hypertension. They might be symptom-free in an early stage, or there may be no symptoms at all during the patient's lifetime. However, the worsening of a cardiomyopathy can cause heart failure (45). Some cardiomyopathies, like the ones caused by the Phospholamban (PLN) p.Arg14del mutation, are rare and need early diagnosis. They are among the most malignant cardiomyopathies and require early implantation of cardioverter-defibrillators (46,47).

Currently, the diagnosis of the PLN mutation is made using genetic testing. These tests, however, are expensive and time-consuming. Some known characteristics of the mutation might be presented in an electrocardiogram (ECG) but, given its rarity, these patterns might not be recognized by a general practitioner at the clinic. Some ML models were able to outperform specialists in the ECG-based detection of PLN mutation**,** nevertheless, the models are still far from a real clinical application (48). These models were trained and evaluated with the same proportion of healthy and PLN patients, which is not correspondent with the rarity of the disease.

Machine Learning

Traditionally, clinical prognostic models are developed using the well-accepted Logistic Regression (LR) technique due to its simplicity and straightforward interpretation. These models usually have a small number of features compared to the number of patients that are included to generate them (49), making them easily interpretable. Their resulting coefficients, representing the extent and direction of the relationship between a feature and the predicted variable, can be used to understand which features are the most impactful in the prediction. LR is usually associated with the traditional (statistical) approach, where the goal is the analysis of the learned coefficients rather than the predictions themselves (49,50). Also, the LR models rely on previous knowledge for appropriate pre-processing to handle non-linear features or multicollinearity (51).

The use of ML has been increasing in current years given the large improvement in computing power and data storage. ML models can deal with non-linearities in the features, combining them in multiple ways automatically (51), which demands computer power. With the advances in deep learning architectures and gradient boosting algorithms, research on the application of ML techniques in the cardiovascular field has significantly increased (52–55).

The use of artificial intelligence in cardiology is seen today as a viable way to optimize the physicians' workload by supporting their decisions and automatizing some of their time-consuming tasks. Deep learning, for instance, is being largely used to perform segmentation or detect diseases on medical images. The models occasionally achieve similar or superior accuracy when compared to physicians (56,57). However, the lack of proper and extensive model validation is still deemed as a limitation (58).

Thesis outline

This thesis presents the development of multiple machine learning models in the field of cardiology, as well as techniques to deal with limitations regarding its implementations and validation. Models were developed for the prediction of TAVI outcomes, recurrence of AF after thoracoscopic surgery, prediction of low attenuation on CTCAs, and detection of a rare genetic disease using only an ECG signal. I also implemented ways to improve the models' accuracy with limited amounts of available data, using interpretation techniques to better understand the models, and validating them with different settings and approaches. My main contributions were: (a) the use of machine learning methods for relevant topics in cardiology, (b) approaches to deal with data sharing policies using finetuning and distributed learning, and (c) optimization and validation of models (internal, external, temporal and subgroup analysis) to analyse model performance and stability, and (d) model explanation techniques for interpretation of the decisions made by the models.

In **Chapter 2** we present the prediction of TAVI outcomes, mortality, and improvement of symptoms using ML techniques. The models were developed using screening and laboratory features, as well as their combination.

We performed an external validation of models trained for the prediction of 1-year mortality after TAVI in **Chapter 3**. To overcome limitations regarding data sharing, the focus was on neural networks with a finetuning strategy, allowing us to take advantage of data from other medical centre to improve the model's accuracy.

Based on the model finetuning strategy described in the previous chapter, in **Chapter 4** we explored distributed and local ML approaches for inter-hospital data utilization. We trained models incrementally in two centres in a distributed way and also used an stacking approach to combine models trained locally on each centre.

Knowing that the TAVI procedure and patient selection changed over time, in **Chapter 5**, the stability and performance of mortality prediction models were analysed over the years. We divided a national registry into temporally organized groups and analysed the data shift and stability of the models over time.

In **Chapter 6** AF is discussed, focusing on the prediction of recurrence after the thoracoscopic surgical ablation. We analysed how available pre-operative features can predict recurrence and in which subgroups the models achieve the highest and lowest accuracy.

Chapter 7 is about the prediction of insufficient attenuation on the ascending aorta to support the diagnosis of coronary artery disease. Patient features, commonly used in contrast protocols, were combined with imaging features extracted from the test bolus contrast.

In **Chapter 8**, an approach to improve the detection of PLN based on ECG is proposed. Also, the models were trained and evaluated in an imbalanced scenario, with a higher proportion of healthy patients over PLN cases, which is closer to clinical practice. The models were pre-trained on an easier task, with larger amounts of ECGs, and later tuned to detect the PLN mutation carriers using the ECG as input.

Finally, in **Chapter 9**, a discussion of the main contributions of this thesis, findings, limitations, clinical implications of the machine learning use in the field of cardiology, and topics for future research is presented.

References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
2. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland Jr JC, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 1—background, design considerations, and model development. Ann Thorac Surg. 2018;105(5):1411–8.
3. Nashef SA, Sharples LD, Roques F, Lockowandt U. EuroSCORE II and the art and science of risk modelling. Eur J Cardiothorac Surg. 2013;43(4):695–6.
4. Bowdish ME, D'Agostino RS, Thourani VH, Desai N, Shahian DM, Fernandez FG, et al. The Society of Thoracic Surgeons Adult Cardiac Surgery Database: 2020 Update on Outcomes and Research. Ann Thorac Surg. 2020 Jun 1;109(6):1646–55.
5. Hickey GL, Grant SW, Bridgewater B. Validation of the EuroSCORE II: should we be concerned with retrospective performance? Eur J Cardio-Thoracic Surg [Internet]. 2013 Mar 1 [cited 2022 Feb 10];43(3):655–655. Available from: https://academic.oup.com/ejcts/article/43/3/655/717862
6. Poullis M, Pullan M, Chalmers J, Mediratta N. The validity of the original EuroSCORE and EuroSCORE II in patients over the age of seventy. Interact Cardiovasc Thorac Surg [Internet]. 2015 Feb 1 [cited 2022 Feb 10];20(2):172–7. Available from: https://pubmed.ncbi.nlm.nih.gov/25348730/
7. Kuplay H, Erdoğan SB, Baştopçu M, Karpuzoğlu E, Er H. Performance of the EuroSCORE II and the STS score for cardiac surgery in octogenarians. Turk gogus kalp damar cerrahisi Derg [Internet]. 2021 [cited 2022 Feb 10];29(2):174–82. Available from: https://pubmed.ncbi.nlm.nih.gov/34104511/
8. Chalmers J, Pullan M, Fabri B, McShane J, Shaw M, Mediratta N, et al. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. Eur J Cardio-Thoracic Surg [Internet]. 2013 Apr 1 [cited 2022 Feb 10];43(4):688–94. Available from: https://academic.oup.com/ejcts/article/43/4/688/441760
9. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785–94.
10. Dorogush AV, Ershov V, Yandex AG. CatBoost: gradient boosting with categorical features support. 2018 Oct 24 [cited 2021 Dec 9]; Available from: https://arxiv.org/abs/1810.11363v1
11. Bazoukis G, Stavrakis S, Zhou J, Bollepalli SC, Tse G, Zhang Q, et al. Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic review. Heart Fail Rev [Internet]. 2021 Jan 1 [cited 2022 Jan 4];26(1):23–34. Available from: https://link.springer.com/article/10.1007/s10741-020-10007-3
12. Benedetto U, Dimagli A, Sinha S, Cocomello L, Gibbison B, Caputo M, et al. Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis. J Thorac Cardiovasc Surg. 2020 Aug 10;
13. Dudchenko A, Ganzinger M, Kopanitsa G. Machine Learning Algorithms in Cardiology Domain: A Systematic Review. Open Bioinforma J [Internet]. 2020

Apr 23 [cited 2022 Jan 4];13(1):25–40. Available from: https://openbioinformaticsjournal.com
14. Nishimura RA. Aortic valve disease. Circulation [Internet]. 2002 Aug 13 [cited 2022 Jan 17];106(7):770–2. Available from: https://www.ahajournals.org/doi/abs/10.1161/01.cir.0000027621.26167.5e
15. Iung B, Baron G, Butchart EG, Delahaye F, Gohlke-Bärwolf C, Levang OW, et al. A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease. Eur Heart J [Internet]. 2003 Jul 1;24(13):1231–43. Available from: https://doi.org/10.1016/S0195-668X(03)00201-X
16. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. Lancet. 2006;368(9540):1005–11.
17. Carabello BA, Paulus WJ. Aortic stenosis. Lancet. 2009;373(9667):956–66.
18. Leon MB, Smith CR, Mack M, Miller DC, Moses JW, Svensson LG, et al. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. N Engl J Med. 2010;363(17):1597–607.
19. Otto CM, Dharam Kumbhani C-CJ, Karen Alexander C-CP, John Calhoon FH, Milind Desai FY, Sanjay Kaul F, et al. 2017 ACC Expert Consensus Decision Pathway for Transcatheter Aortic Valve Replacement in the Management of Adults With Aortic Stenosis: A Report of the American College of Cardiology Task Force on Clinical Expert Consensus Documents. J Am Coll Cardiol [Internet]. 2017 Mar 14 [cited 2022 Jan 3];69(10):1313–46. Available from: http://www.elsevier.com/about/policies/author-agreement/obtaining-permission
20. Fang F, Tang J, Zhao Y, He J, Xu P, Faramand A. Transcatheter aortic valve implantation versus surgical aortic valve replacement in patients at low and intermediate risk: A risk specific meta-analysis of randomized controlled trials. PLoS One [Internet]. 2019 Sep 24;14(9):e0221922-. Available from: https://doi.org/10.1371/journal.pone.0221922
21. Khan AA, Murtaza G, Khalid MF, Khattak F. Risk Stratification for Transcatheter Aortic Valve Replacement. Cardiol Res [Internet]. 2019 [cited 2021 Dec 9];10(6):323. Available from: /pmc/articles/PMC6879047/
22. Edwards FH, Cohen DJ, O'Brien SM, Peterson ED, Mack MJ, Shahian DM, et al. Development and Validation of a Risk Prediction Model for In-Hospital Mortality After Transcatheter Aortic Valve Replacement. JAMA Cardiol [Internet]. 2016 Apr 1 [cited 2021 Dec 9];1(1):46–52. Available from: https://pubmed.ncbi.nlm.nih.gov/27437653/
23. Iung B, Laouénan C, Himbert D, Eltchaninoff H, Chevreul K, Donzeau-Gouge P, et al. Predictive factors of early mortality after transcatheter aortic valve implantation: individual risk assessment using a simple score. Heart [Internet]. 2014 Jul 1 [cited 2021 Dec 9];100(13):1016–23. Available from: https://heart.bmj.com/content/100/13/1016
24. Al-Farra H, Abu-Hanna A, de Mol BAJM, ter Burg WJ, Houterman S, Henriques JPS, et al. External validation of existing prediction models of 30-day mortality after Transcatheter Aortic Valve Implantation (TAVI) in the Netherlands Heart Registration. Int J Cardiol. 2020 Oct 15;317:25–32.
25. Silva LS, Caramori PRA, Nunes Filho ACB, Katz M, Guaragna JCV da C, Lemos P, et al. Performance of Surgical Risk Scores to Predict Mortality after

Transcatheter Aortic Valve Implantation. Arq Bras Cardiol [Internet]. 2015 Jul 31 [cited 2021 Dec 9];105(3):241–7. Available from: http://www.scielo.br/j/abc/a/XjmTNY6kkJsbkM95dnhsDkm/?lang=en
26. Neumann FJ, Sechtem U, Banning AP, Bonaros N, Bueno H, Bugiardini R, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J [Internet]. 2020 Jan 14 [cited 2022 Mar 4];41(3):407–77. Available from: https://pubmed.ncbi.nlm.nih.gov/31504439/
27. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: A report from the American Heart Association. Circulation. 2020;E139–596.
28. Timmis A, Townsend N, Gale CP, Torbica A, Lettino M, Petersen SE, et al. European Society of Cardiology: Cardiovascular Disease Statistics 2019. Eur Heart J [Internet]. 2020 Jan 1 [cited 2022 Mar 7];41(1):12–85. Available from: https://academic.oup.com/eurheartj/article/41/1/12/5670482
29. Marano R, Rovere G, Savino G, Flammia FC, Carafa MRP, Steri L, et al. CCTA in the diagnosis of coronary artery disease. Radiol Med [Internet]. 2020;125(11):1102–13. Available from: https://doi.org/10.1007/s11547-020-01283-y
30. W. SP, Hironori H, Scot G, Hideyuki K, L. NB, R. DM, et al. Coronary Computed Tomographic Angiography for Complete Assessment of Coronary Artery Disease. J Am Coll Cardiol [Internet]. 2021 Aug 17;78(7):713–36. Available from: https://doi.org/10.1016/j.jacc.2021.06.019
31. Isogai T, Jinzaki M, Tanami Y, Kusuzaki H, Yamada M, Kuribayashi S. Body weight-tailored contrast material injection protocol for 64-detector row computed tomography coronary angiography. Jpn J Radiol. 2011;29(1):33–8.
32. van den Boogert TPW, Lopes RR, Lobe NHJ, Verwest TA, Stoker J, Henriques JP, et al. Patient-tailored Contrast Delivery Protocols for Computed Tomography Coronary Angiography: Lower Contrast Dose and Better Image Quality. J Thorac Imaging. 2021 Jul;
33. Alizadehsani R, Abdar M, Roshanzamir M, Khosravi A, Kebria PM, Khozeimeh F, et al. Machine learning-based coronary artery disease diagnosis: A comprehensive review. Comput Biol Med. 2019 Aug 1;111:103346.
34. Arrhythmia | NHLBI, NIH [Internet]. [cited 2022 Jan 3]. Available from: https://www.nhlbi.nih.gov/health-topics/arrhythmia
35. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. Circulation. 2019 Mar 5;139(10):e56–528.
36. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. Eur J cardio-thoracic Surg. 2016;50(5):e1–88.
37. Hindricks G, Potpara T, Dagres N, Bax JJ, Boriani G, Dan GA, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Vol. 42, European Heart Journal. Oxford University Press; 2021. p. 373–498.
38. Boersma LVA, Castella M, van Boven W, Berruezo A, Yilmaz A, Nadal M, et al. Atrial fibrillation catheter ablation versus surgical ablation treatment (FAST) a 2-center randomized clinical trial. Circulation. 2012;125(1):23–30.

39. Driessen AHG, Berger WR, Krul SPJ, van den Berg NWE, Neefs J, Piersma FR, et al. Ganglion Plexus Ablation in Advanced Atrial Fibrillation. J Am Coll Cardiol. 2016 Sep;68(11):1155–65.
40. Rizwan A, Zoha A, Mabrouk I Ben, Sabbour HM, Al-Sumaiti AS, Alomainy A, et al. A Review on the State of the Art in Atrial Fibrillation Detection Enabled by Machine Learning. IEEE Rev Biomed Eng. 2021;14:219–39.
41. Matias I, Garcia N, Pirbhulal S, Felizardo V, Pombo N, Zacarias H, et al. Prediction of Atrial Fibrillation using artificial intelligence on Electrocardiograms: A systematic review. Comput Sci Rev. 2021 Feb 1;39:100334.
42. Vitali F, Serenelli M, Airaksinen J, Pavasini R, Tomaszuk-Kazberuk A, Mlodawska E, et al. CHA2DS2-VASc score predicts atrial fibrillation recurrence after cardioversion: Systematic review and individual patient pooled meta-analysis. Clin Cardiol [Internet]. 2019 Mar 1 [cited 2022 Jan 16];42(3):358. Available from: /pmc/articles/PMC6712331/
43. Kosich F, Schumacher K, Potpara T, Lip GY, Hindricks G, Kornej J. Clinical scores used for the prediction of negative events in patients undergoing catheter ablation for atrial fibrillation. Clin Cardiol. 2019;42(2):320–9.
44. Deng H, Bai Y, Shantsila A, Fauchier L, Potpara TS, Lip GYH. Clinical scores for outcomes of rhythm control or arrhythmia progression in patients with atrial fibrillation: a systematic review. Clin Res Cardiol. 2017;106(10):813–23.
45. Wexler RK, Elton T, Pleister A, Feldman D. Cardiomyopathy: An Overview. Am Fam Physician [Internet]. 2009 [cited 2022 Jan 4];79(9):778. Available from: /pmc/articles/PMC2999879/
46. Bosman LP, Verstraelen TE, van Lint FHM, Cox MGPJ, Groeneweg JA, Mast TP, et al. The Netherlands Arrhythmogenic Cardiomyopathy Registry: design and status update. Neth Heart J [Internet]. 2019 Oct 1 [cited 2022 Jan 4];27(10):480–6. Available from: https://pubmed.ncbi.nlm.nih.gov/30997596/
47. Van Rijsingen IAW, Van Der Zwaag PA, Groeneweg JA, Nannenberg EA, Jongbloed JDH, Zwinderman AH, et al. Outcome in phospholamban R14del carriers results of a large multicentre cohort study. Circ Cardiovasc Genet [Internet]. 2014 Aug 1 [cited 2022 Jan 4];7(4):455–65. Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCGENETICS.113.000374
48. Bleijendaal H, Ramos LA, Lopes RR, Verstraelen TE, Baalman SWE, Oudkerk Pool MD, et al. Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p.Arg14del mutation on the electrocardiogram? Hear Rhythm. 2021 Jan 1;18(1):79–87.
49. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics [Internet]. 2018 Jul 17 [cited 2022 Jan 13];19(1):1–14. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5
50. Menard S. Applied logistic regression analysis. Vol. 106. Sage; 2002.
51. Levy JJ, Levy JJ, Levy JJ, O'Malley AJ, O'Malley AJ. Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning. BMC Med Res Methodol [Internet]. 2020 Jun 29 [cited 2022 Jan 13];20(1):1–15. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01046-3

52. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial Intelligence in Cardiology: Present and Future. Mayo Clin Proc [Internet]. 2020 May 1 [cited 2022 Jan 13];95(5):1015–39. Available from: https://pubmed.ncbi.nlm.nih.gov/32370835/
53. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol [Internet]. 2018 Jun 12 [cited 2022 Jan 13];71(23):2668–79. Available from: https://pubmed.ncbi.nlm.nih.gov/29880128/
54. Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Health Care. Ann Thorac Surg [Internet]. 2020 May 1 [cited 2022 Jan 13];109(5):1323–9. Available from: https://pubmed.ncbi.nlm.nih.gov/31706869/
55. Friedrich S, Groß S, Kö Nig IR, Engelhardt S, Bahls M, Heinz J, et al. Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: a systematic review with recommendations. Eur Hear J - Digit Heal [Internet]. 2021 Sep 30 [cited 2022 Jan 13];2(3):424–36. Available from: https://academic.oup.com/ehjdh/article/2/3/424/6294933
56. Liu X, Song L, Liu S, Zhang Y. A Review of Deep-Learning-Based Medical Image Segmentation Methods. Sustain 2021, Vol 13, Page 1224 [Internet]. 2021 Jan 25 [cited 2022 Jul 10];13(3):1224. Available from: https://www.mdpi.com/2071-1050/13/3/1224/htm
57. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal. 2019 Oct 1;1(6):e271–97.
58. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ [Internet]. 2020 Mar 25 [cited 2022 Jul 10];368. Available from: https://www.bmj.com/content/368/bmj.m689

2

Value of machine learning in predicting TAVI outcomes

Lopes RR, van Mourik MS, Schaft EV, Ramos LA, Baan Jr J, Vendrik J, de Mol BA, Vis MM, Marquering HA.

Netherlands Heart Journal. 2019 Sep;27(9):443-50.

DOI: 10.1007/s12471-019-1285-7

Abstract

Transcatheter aortic valve implantation (TAVI) has become a commonly applied procedure for high-risk aortic valve stenosis patients. However, for some patients, this procedure does not result in the expected benefits. Previous studies indicated that it is difficult to predict the beneficial effects for specific patients. We aim to study the accuracy of various traditional machine learning (ML) algorithms in the prediction of TAVI outcomes.

Clinical and laboratory data from 1,478 TAVI patients from a single centre were collected. The outcome measures were improvement of dyspnoea and mortality. Three experiments were performed using (1) screening data, (2) laboratory data, and (3) the combination of both. Five well-established ML techniques were implemented, and the models were evaluated based on the area under the curve (AUC). Random forest classifier achieved the highest AUC (0.70) for predicting mortality. Logistic regression had the highest AUC (0.56) in predicting improvement of dyspnoea.

In our single-centre TAVI population, the tree-based models were slightly more accurate than others in predicting mortality. However, ML models performed poorly in predicting improvement of dyspnoea.

Introduction

Aortic valve stenosis (AS) is one of the most common valvular heart diseases, impacting, in general, the elderly population. In the past decade, transcatheter aortic valve implantation (TAVI) has developed into a routine treatment for AS patients at elevated risk of surgery. Although there is strict patient selection for the TAVI procedure and various planning and treatment support tools are available (1–3), a number of patients have limited benefit from TAVI (4). Improved selection of these patients would allow increased benefit from the procedure and improve decision-making. Unfortunately, current risk models have only limited accuracy in predicting TAVI outcomes (5).

Previous clinical prediction models rely on traditional statistical regression models (6). Alternatively, machine learning (ML), which is a computer science subdiscipline, has shown superior predictive value in various clinical areas, from detecting Alzheimer's disease to identifying lung nodules (7,8). A more specific area of ML is supervised learning: with known outcomes, ML algorithms can learn automatically to optimise the prediction of this outcome. Moreover, ML techniques have outperformed conventional regression models when applied to a large amount of data (9).

Multiple risk models that have been used that are dedicated to the prediction of perioperative mortality and are not TAV-specific, but intended for surgical aortic valve replacement such as the EuroSCORE, EuroSCORE II or the STS (Society of Thoracic Surgery) score (10,11). For TAVI, these are poor predictors of mortality and focus on procedural or 30-day mortality, as did the TAVI-specific TVT registry score (12). The prediction of 1-year mortality is even more challenging (13). A more recent study also incorporated predefined features from computed tomography (CT) in combination with comorbidities to enhance the model (14).

We aimed to study the accuracy of various ML algorithms in predicting outcomes after a TAVI procedure. The accuracy was evaluated in the prediction of mortality and improvement of dyspnoea using a subset of well-established ML techniques.

Methods

Patient population

The database consists of 1,478 patients who underwent a TAVI between 2007 and 2018; their median age is 82.9 years [Q_1 78.0 – Q_3 86.4] and 55% of the patients are female. The data contain patient characteristics, medical history, symptoms, and test results prior to and after TAVI. Symptoms are dyspnoea, fatigue, collapse, and angina pectoris. Tests performed prior to TAVI are echocardiography, computed tomography angiography, coronary angiography, electrocardiography (ECG), and laboratory tests. Tests done after TAVI are echocardiography, ECG and laboratory tests.

The outcomes used are improvement of dyspnoea and mortality. Dyspnoea is measured using the New York Heart Association (NYHA) functional score (1-4). Mortality is defined as a patient who died of a cardiovascular disease within 1 year after the procedure. Patients with missing data are excluded. The baseline and 60-day follow-up NYHA score is known for 766 patients (605 improved, 161 non-improvements) and mortality is known for 1,400 patients (1,263 survivors, 137 non-survivors). For every outcome parameter, we performed three experiments: (1) using only screening data; (2) using only laboratory data; and (3) using both screening and laboratory data. The number of patients for each experiment is different due to missing values, as presented in Figures 1 and 2. All variables, as well as the descriptive statistics, can be found in the Supplementary Material (Tables I and II).

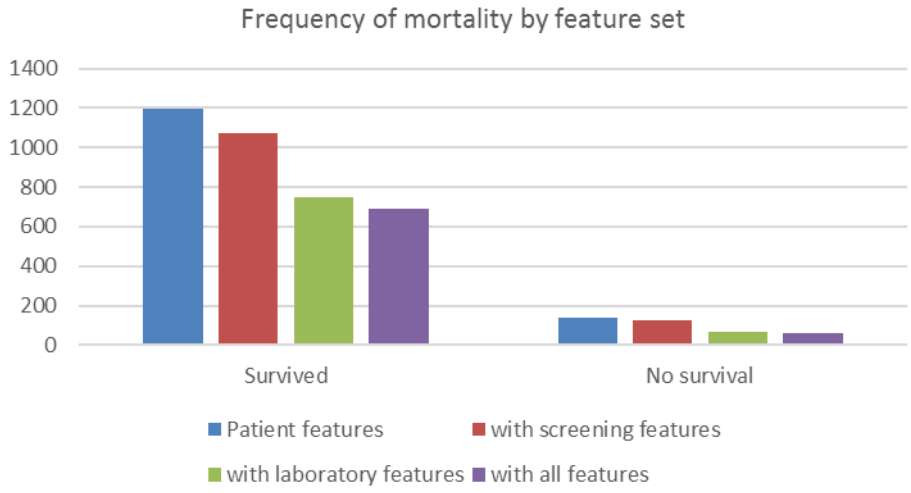

Figure 1. Number of patients without missing data per feature set for mortality outcome. For each feature set added, a lower number of samples is available due to missing values in different patients per set.

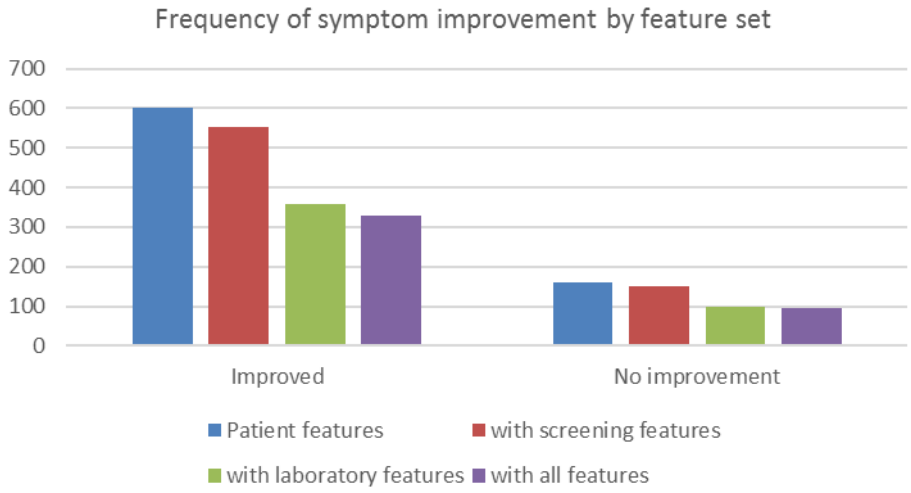

Figure 2. Number of patients without missing data per feature set for symptom outcome. For each feature set added, a lower number of samples is available due to missing values in different patients per set.

Clinical variables

The clinical variables used can be divided into three sets: patient characteristics, screening data, and laboratory data. The patient baseline characteristics included

age, sex, body mass index (BMI), and access route chosen for the procedure. The screening data consist of the medical history, symptoms, and echocardiography prior to TAVI. The features used from this data are the existence of peripheral artery disease, chronic obstructive pulmonary disease (COPD), atrial fibrillation, diabetes mellitus (DM) in the medical history, left ventricular function, and aortic valve area assessed with echocardiography. The features used from the laboratory data are the pre-procedural values of N-terminal pro-b-type natriuretic peptide (NT-proBNP), haemoglobin, albumin, chronic kidney disease epidemiology collaboration (CKD-EPI), and creatinine.

Based upon expert opinions we applied clipping to the NT-proBNP and creatinine variables, for values greater than 1,000 ng/l and 250 mmol/l, respectively. The nominal and categorical data were one-hot encoded; continuous features were normalised by removing the mean and scaling to unit variance, as requisite for many ML techniques (15). Moreover, the COPD and DM were dichotomised to take into account the presence of the disease instead of the degree.

Classification techniques

In this study, we selected a number of well-established ML techniques, which are: support vector machine (SVM) (16), random forest classifier (RFC) (17), multi-layer perceptron (ML) (18), and gradient tree boosting (GTB) (19). In addition, traditional logistic regression (LR) was also applied for comparison, since this technique is often used in clinical studies. All the implementations used in this project were provided by scikit-learn (15), except for GTB. We chose the XGBoost (19) library because of its GPU implementation, which speeds up training and optimisation.

To evaluate the models fairly, the database was split into two sets: a training and a testing dataset. The training data were used to find the optimal parameters for the classification task. The testing set was used to evaluate the trained model in unseen data, to ensure generalisation of the model and prevent the memorisation of the training set (overfitting). In this study, the models were evaluated with the Monte-Carlo cross-validation for 100 iterations and stratified splits of 70% for training and 30% for testing. With this large number of different training and testing sets, chances of having over-optimistic results are minimised. Moreover,

to optimise the parameters of each model, a randomised grid search with stratified 5-fold cross-validation was performed using the training set. The hyperparameters and ranges used for optimisation, including the weight penalisation applied to minimise the class unbalance issue, are available in the Supplementary Material (Tables III and IV).

Results of ML techniques are difficult to interpret. To elucidate which features may be important in the ML techniques, the average feature importance for RFC and GTB was calculated based on the number of times the feature was selected for splitting and weighted by the average squared improvement of the model over all trees (20).

Performance assessment

The median of the area under the curve (AUC) of the receiver operating characteristic curve (ROC) from 100 iterations, using test sets, was selected to evaluate the performance of each model. To assess whether the difference in AUC between highest performing classifier and the other methods was statistically significant, the Wilcoxon signed-rank test was performed for each experiment. *p*-values < 0.05 were considered statistically significant.

Results

The predictive value for improvement of dyspnoea was statistically significant but absent/low, with the best median AUC result of 0.56, using only laboratory features and LR. For mortality prediction, the model based on RFC was most accurate with an AUC of 0.70 [Q_1 0.67 – Q_3 0.74] and the results are considered to be significantly different according to the Wilcoxon test. All results are presented in Table 1. There was no significant difference in the results for 1-year mortality prediction using only the screening features.

The combination of the feature data sets did not result in an increased AUC in predicting improvement of dyspnoea. In mortality prediction, the models using the data combination showed similar AUCs using only laboratory features and all features. The median receiver operating characteristic (ROC) curves for the prediction of dyspnoea improvement (using the laboratory features) and mortality prediction (using all features) are displayed in Figures 3 and 4, respectively.

Table 1. Median area under the curve [first and third quartiles] for all experiments. The *rows* are the machine learning technique and the *columns* are the set of features and the kind of outcome prediction. The highest-performing models and the models proved to be insignificantly different from those according to the Wilcoxon test are highlighted in *bold*.

	Improvement of dyspnoea			1-year mortality		
Model	***Screening***	***Laboratory***	***All***	***Screening***	***Laboratory***	***All***
GTB	0.52 [0.49-0.56]	0.53 [0.50-0.55]	0.51 [0.47-0.54]	0.65 [0.62-0.67]	0.69 [0.65-0.72]	0.69 [0.66-0.72]
SVM	0.52 [0.49-0.55]	0.52 [0.48-0.56]	0.53 [0.48-0.56]	0.65 [0.62-0.68]	0.68 [0.64-0.71]	0.69 [0.65-0.72]
MLP	0.53 [0.50-0.56]	0.52 [0.48-0.55]	0.52 [0.48-0.56]	0.65 [0.62-0.68]	0.66 [0.62-0.70]	0.66 [0.62-0.71]
RFC	0.52 [0.49-0.55]	0.53 [0.49-0.56]	0.51 [0.46-0.56]	0.66 [0.63-0.68]	0.70 [0.67-0.73]	0.70 [0.67-0.74]
LR	0.54 [0.52-0.57]	0.56 [0.52-0.58]	0.54 [0.51-0.57]	0.66 [0.63-0.69]	0.67 [0.62-0.70]	0.65 [0.61-0.69]

GTB gradient tree boosting, *SVM* support vector machine, *MLP* multi-layer perceptron, *RFC* random forest classifier, *LR* logistic regression

The most relevant features for mortality prediction in GTB and RFC were determined by the importance of the features. In order of relevance, these were: NT-proBNP, BMI , CKD-EPI, creatinine, and age.

Discussion

In our population of 1,478 patients who underwent a TAVI procedure, the selected subset of ML techniques had little added prognostic value in predicting mortality and improvement of dyspnoea compared to commonly applied LR techniques. In the prediction of mortality, ML techniques achieved similar scores using all features and only the laboratory features. The increase in prognostic value and the improvement of dyspnoea prediction was rather low, even with the combination of clinical and laboratory data.

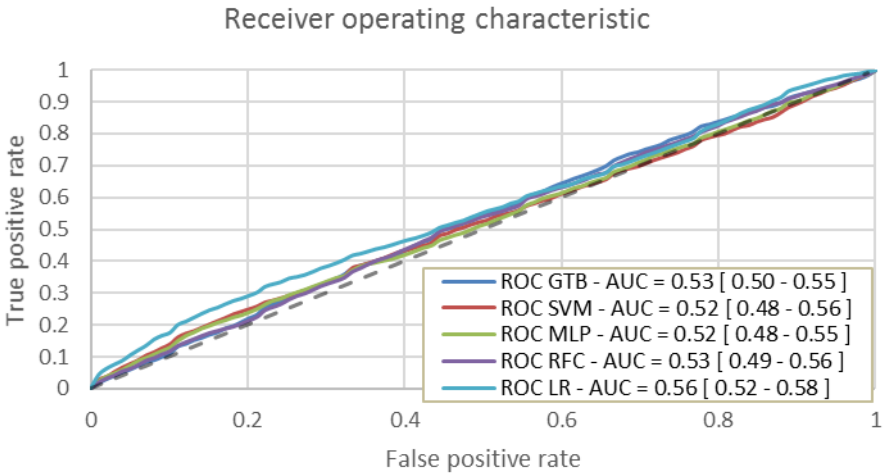

Figure 3. Median receiver operating characteristic (ROC) curve from 100 Monte Carlo cross-validation iterations for the prediction of dyspnoea improvement using laboratory features. *AUC* area under the curve, *GTB* gradient tree boosting, *LR* logistic regression, *MLP* multi-layer perceptron, *RFC* random forest classifier, *SVM* support vector machine

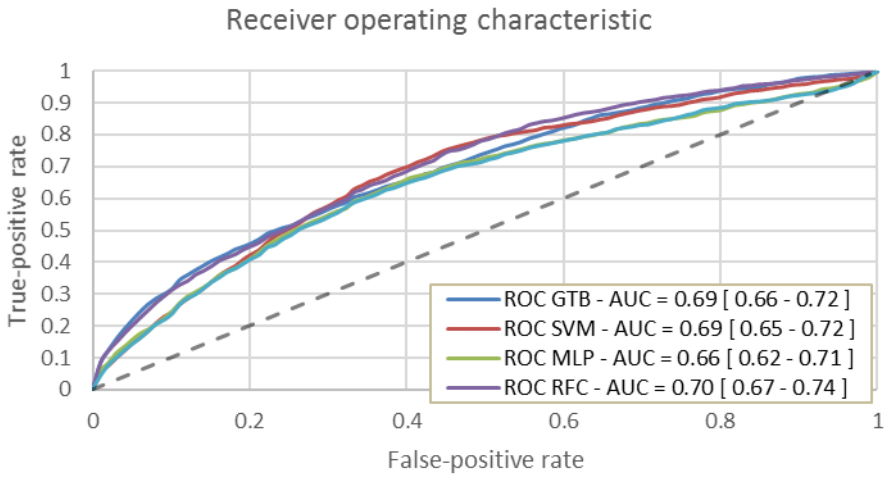

Figure 4. Median ROC curve from 100 Monte Carlo cross-validation iterations for the mortality prediction using all features. *AUC* area under the curve, *GTB* gradient tree boosting, *LR* logistic regression, *MLP* multi-layer perceptron, *RFC* random forest classifier, *SVM* support vector machine

Some recent studies that applied ML have often shown positive results for prognosis prediction. In the study of Memarian et al. (21), ML methods were applied to multimodal data (clinical data, electroencephalography, magnetic resonance imaging) to predict the outcome of surgery in patients with mesial temporal lobe epilepsy, achieving a prediction accuracy of 95% using SVM-derived classifiers. Frizzell et al. (22) compared ML methods to LR in predicting 30-day readmission in patients discharged following hospitalisation for heart failure. Similar to our findings these results did not show an improvement in prediction accuracy. The prediction of six cardiovascular outcomes (including heart failure and all-cause death) was assessed by Ambale-Venkatesh et al. (23), whereby random survival forests and other ML techniques were compared to the standard cardiovascular scores. They concluded that ML improved the accuracy of cardiovascular event prediction in initially asymptomatic patients.

Our results confirm that predicting outcomes of TAVI procedures is challenging. Many factors may impact the patient's outcome, many of which are not considered in the modelling. The inclusion of more and different kinds of features, such as different examinations, CT scans, and ECG, is currently a subject of investigation. By including different sets of features and more complex models, the predictive value may increase.

There was no implicit order in the data variables that we tried to exploit. Also, no variables were transformed into a dense representation. We included all variables that were considered relevant by clinical experts. We are aware that one-hot encoding generates data sparsity. Even though one-hot encoding can downgrade the performance of some ML methods, it is an important step for distance-based methods such as the SVM. In our study, only a small number of categorical variables (with few classes) were one-hot encoded to prevent hampering the performance of methods due to data sparsity.

The methods used in this study are generalisable to other clinical challenges in which prediction of outcomes is warranted. It is expected that the application of ML techniques in combination with clinical knowledge will become increasingly important in coming years to improve prognostics. Models with higher accuracy may improve outcome prediction after TAVI, allowing a more individual approach in clinical care.

This study suffered from a number of limitations. The dataset used in this study may be one of the largest Dutch single-centre TAVI datasets available. However, with unbalanced measures (such as a relatively small population that did not survive the 1st year), the effect of the data is reduced. Moreover, many patients were excluded because of missing data, which can be mitigated by using imputation techniques. In this study, we chose symptom reduction using the NYHA classification and 1-year mortality as outcome measures. Other outcome measures, however, might be relevant for the TAVI population.

Conclusion

In our population of patients treated with TAVI, ML techniques were able to predict mortality using the current set of features. In predicting a reduction of dyspnoea, the traditional LR technique outperformed the others. Adding more features or increasing the dataset size may result in a situation in which ML techniques have more added value.

References

1. Khalil A, Faisal A, Lai KW, Ng SC, Liew YM. 2D to 3D fusion of echocardiography and cardiac CT for TAVR and TAVI image guidance. Med Biol Eng Comput. 2017;55(8):1317–26.
2. Grbic S, Mansi T, Ionasec R, Voigt I, Houle H, John M, et al. Image-based computational models for TAVI planning: From CT images to implant deployment. Med Image Comput Comput Interv. 2013;395–402.
3. Swee JKY, Grbić S. Implantation (TAVI) Planning from CT with ShapeForest. Med Image Comput Comput Interv – MICCAI. 2014;17–24.
4. Puri R, Iung B, Cohen DJ, Rodés-Cabau J. TAVI or No TAVI: Identifying patients unlikely to benefit from transcatheter aortic valve implantation. Eur Heart J. 2016;37(28):2217–25.
5. Martin GP, Sperrin M, Ludman PF, de Belder MA, Gale CP, Toff WD, et al. Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation. Am Heart J. 2017;184:97–105.
6. Van Mourik MS, Vendrik J, Abdelghani M, Van Kesteren F, Henriques JPS, Driessen AHG, et al. Guideline-defined futility or patient-reported outcomes to assess treatment success after TAVI: What to use? Results from a prospective cohort study with long-term follow-up. Open Hear. 2018;5(2).
7. Lebedev A V., Westman E, Van Westen GJP, Kramberger MG, Lundervold A, Aarsland D, et al. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. NeuroImage Clin. 2014;6:115–25.
8. Nishio M, Nishizawa M, Sugiyama O, Kojima R, Yakami M, Kuroda T, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. PLoS One. 2018;13(4):1–13.
9. Singal A., Mukherjee A., Elmunzer B.J., Higgins P.D., Lok A.S., Zhu J., et al. Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma. Am J Gastroenterol. 2016;42(2):407–20.
10. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 2-Isolated Valve Surgery. Ann Thorac Surg [Internet]. 2009;88(1 SUPPL.):S23–42. Available from: http://dx.doi.org/10.1016/j.athoracsur.2009.05.056
11. Lemeshow S, Gauducheau E, Roques F, Nashef SAM, Michel P, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardio-Thoracic Surg. 2002;16(1):9–13.
12. Pilgrim T, Franzone A, Stortecky S, Nietlispach F, Haynes AG, Tueller D, et al. Predicting Mortality After Transcatheter Aortic Valve Replacement: External Validation of the Transcatheter Valve Therapy Registry Model. Circ Cardiovasc Interv. 2017;10(11):1–9.
13. Ludman PF, Moat N, De Belder MA, Blackman DJ, Duncan A, Banya W, et al. Transcatheter aortic valve implantation in the United Kingdom: Temporal trends, predictors of outcome, and 6-year follow-up: A report from the UK transcatheter aortic valve implantation (TAVI) registry, 2007 to 2012. Circulation. 2015;131(13):1181–90.

14. Lantelme P, Eltchaninoff H, Rabilloud M, Souteyrand G, Dupré M, Spaziano M, et al. Development of a Risk Score Based on Aortic Calcification to Predict 1-year Mortality After Transcatheter Aortic Valve Replacement. JACC Cardiovasc Imaging. 2018;
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012;12:2825–30.
16. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273–97.
17. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
18. Bishop CM. Neural Networks for Pattern Recognition. Oxford Univ Press. 1995;
19. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;785–94.
20. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. Stat Med. 2003;22(9):1365–81.
21. Memarian N, Kim S, Dewar S, Engel J, Staba RJ. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. Comput Biol Med. 2015;64:67–78.
22. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. JAMA Cardiol. 2017;2(2):204–9.
23. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Gregory Hundley W, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017;121(9):1092–101.

Supplementary material

Table I. Summarized patient characteristics grouped by symptoms

			Grouped by symptoms	
		Missing	**Improved**	**Didn't improve**
variable	**level**			
n			605	161
Gender	*Female*	36	334 (55.39)	98 (60.87)
	Male		269 (44.61)	63 (39.13)
Body Mass Index		141	27.10 [24.50,30.40]	26.40 [23.70,30.50]
Age (years)		143	83.40 [77.90,86.60]	82.70 [77.50,86.10]
PAD	*No*	155	447 (74.13)	112 (69.57)
	Yes		156 (25.87)	49 (30.43)
COPD	*No*	156	414 (68.66)	114 (70.81)
	Yes		189 (31.34)	47 (29.19)
Atrial Fibrillation	*No*	171	350 (58.04)	109 (67.7)
	Yes, unknown type		77 (12.77)	14 (8.7)
	Yes, paroxysmal		84 (13.93)	17 (10.56)
	Yes, permanent		78 (12.94)	15 (9.32)
	Yes, persistent		14 (2.32)	6 (3.73)
Diabetes mellitus	*No*	166	417 (69.15)	110 (68.32)
	Yes		186 (30.85)	51 (31.68)
Treatment for Diabetes	*Diet alone*	92	2 (0.33)	
	Insulin		33 (5.46)	5 (3.11)
	No		461 (76.32)	129 (80.12)
	Oral medication		101 (16.72)	23 (14.29)
	Oral medication and insulin		7 (1.16)	4 (2.48)
Left Ventricular Function	*Good*	147	371 (61.63)	108 (67.08)

	Mildly impaired		110 (18.27)	31 (19.25)
	Moderately impaired		74 (12.29)	11 (6.83)
	Poor		39 (6.48)	11 (6.83)
	Very poor		8 (1.33)	
Aortic Valve Area (cm^2)		262	0.80 [0.70,1.00]	0.81 [0.70,0.95]
NT-proBNP (ng/L)		448	1644 [641,3957]	1258 [595,3373]
Hemoglobin (mmol/L)		170	7.79 (1.01)	7.85 (1.05)
Albumin (g/L)		453	42 [40,44]	42.00 [39,45]
CKD-EPI ($ml/min/1.73\ m^2$)		199	59.89 [46.03,75.27]	59.00 [43.51,73.44]
Creatinine (mmol/L)		174	88 [73,112]	89 [71,113]
Access Route	*Direct aorta*	162	72 (11.94)	32 (20.0)
	Transapical		57 (9.45)	21 (13.12)
	Transfemoral		474 (78.61)	107 (66.88)
	Via arteria subclavia			0 (0.00)

Table II. Summarized patient characteristics grouped by 1-year mortality

			Grouped by 1-year mortality	
		Missing	**Survived**	**Didn't survive**
variable	**level**			
n			1263	137
Gender	*Female*	36	696 (55.77)	67 (48.91)
	Male		552 (44.23)	70 (51.09)
Body Mass Index		141	27.00 [24.20,30.40]	25.50 [23.70,29.10]
Age (years)		143	82.80 [77.80,86.20]	84.00 [80.00,87.40]
PAD	*No*	155	909 (75.19)	84 (61.31)
	Yes		300 (24.81)	53 (38.69)
COPD	*No*	156	871 (72.16)	74 (54.01)

	Yes		336 (27.84)	63 (45.99)
	No	171	734 (61.58)	70 (51.09)
	Yes, unknown type		127 (10.65)	28 (20.44)
Atrial Fibrillation	*Yes, paroxysmal*		161 (13.51)	17 (12.41)
	Yes, permanent		137 (11.49)	20 (14.6)
	Yes, persistent		33 (2.77)	2 (1.46)
	No	166	833 (69.47)	92 (67.15)
Diabetes mellitus	*Yes*		366 (30.53)	45 (32.85)
	Diet alone	92	10 (0.8)	
	Insulin		53 (4.24)	7 (5.11)
Treatment for Diabetes	*No*		958 (76.7)	104 (75.91)
	Oral medication		200 (16.01)	23 (16.79)
	Oral medication and insulin		28 (2.24)	3 (2.19)
	Good	147	756 (62.84)	59 (43.38)
	Mildly impaired		227 (18.87)	32 (23.53)
Left Ventricular Function	*Moderately impaired*		138 (11.47)	26 (19.12)
	Poor		71 (5.9)	16 (11.76)
	Very poor		11 (0.91)	3 (2.21)
Aortic Valve Area (cm^2)		262	0.80 [0.66,0.95]	0.80 [0.66,0.97]
NT-proBNP (ng/L)		448	1412 [569,3332]	3365 [1541,7007]
Hemoglobin (mmol/L)		170	7.80 (1.00)	7.65 (1.17)
Albumin (g/L)		453	42 [40,44]	41 [38,43]
CKD-EPI (ml/min/1.73 m^2)		199	60.10 [45.90,74.54]	48.27 [33.66,65.41]
Creatinine (mmol/L)		174	88 [72,111]	102 [78,154]
	Direct aorta	162	214 (17.58)	35 (25.55)
Access Route	*Transapical*		109 (8.96)	23 (16.79)
	Transfemoral		894 (73.46)	78 (56.93)

	Via arteria subclavia	1 (0.73)

Table III. Hyperparameters used for SVM

Classifier	Kernel Type	Penalty parameter *C*	Kernel coefficient γ	Degree of the Polynomial kernel	Class weight
SVM	Linear	[0.001, 0.01, 0.1, 1, 10, 100]	n.a.	n.a.	[balanced, 1:5, 1:7, 1:10, 1:13, 1:15]
	Radial basis function	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	n.a.	[balanced, 1:5, 1:7, 1:10, 1:13, 1:15]
	Polynomial	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	[1, 2, 3, 4, 5, 6]	[balanced, 1:5, 1:7, 1:10, 1:13, 1:15]
	Sigmoid	[0.001, 0.01, 0.1, 1, 10, 100]	[1, 0.1, 0.01, 0.001, 0.0001]	n.a.	[balanced, 1:5, 1:7, 1:10, 1:13, 1:15]

Table IV. Hyperparameters used for RFC, MLP and GTB

Classifier	Parameter Name	Parameter Value
RFC	*Number of trees*	[10, 20, 50, 100, 400, 800, 1200, 1600, 2000]
	Max features for split	None, auto, sqrt, log2
	Quality of split	Gini, entropy
	Max depth	[None, 3, 5, 7, 9, 11, 13, 20, 50]
	Min samples per split	2, 4, 6, 8, 10, 20
	Min samples per leaf	2, 4, 6, 8, 10, 20
	Class weight	[balanced, 1:5, 1:7, 1:10, 1:13, 1:15]
MLP	*Hidden Layer sizes*	[4, 4], [4, 8, 4], [50, 25], [50, 25, 10], [70, 40, 20], [70, 30], [50, 30, 20, 10]
	Regularization parameter	[0.1, 0.01, 0.001, 0.0001]
	Batch size	[8, 16, 32, 64]

	Learning rate	[0.01, 0.001]
	Activations	[Relu, Logistic, TanH]
	Optimization	Adam, L-BFGS
GTB	*Minimum child weight*	[1, 5, 10, 15]
	Gamma	[0, 1, 3, 5]
	Subsample ratio	[0.6, 0.8, 1.0]
	Subsample ratio of columns	[0.6, 0.8, 1.0]
	Max depth	[4, 7, 10, 15]
	Class weight	[5, 7, 10, 13, 15]
	Max delta step	[0, 1, 5, 10]
	Number of trees	[30, 50, 100, 200]

3

Inter-center cross-validation and finetuning without patient data sharing for predicting transcatheter aortic valve implantation outcome

Lopes RR*, Mamprin M*, Zelis JM, Tonino PA, van Mourik MS, Vis MM, Zinger S, de Mol BA, de With PH, Marquering HA.

In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) 2020 Jul 28 (pp. 591-596). IEEE.

* Shared first author

DOI: 10.1109/CBMS49503.2020.00117

Abstract

Transcatheter aortic valve implantation (TAVI) is the routine treatment worldwide for aortic valve stenosis in low- to high-risk patients. Assessing patient risk is essential to identify the most suitable candidates that could benefit from the procedure. Despite the broad use of statistical predictors in patient selection, current machine learning predictors have only been validated on retrospective data collected in single centers. Further, external validation is needed to assess the improvement in accuracy, which is offered by machine learning and deep learning techniques. In this study, we propose a finetuning approach for deep learning models by performing an inter-center cross-validation and finetuning technique, in order to improve the cross-validation accuracy results. We aimed to overcome data exchange and policy-related issues of two medical centers with a dedicated protocol, exploiting the exchange of deep learning models, data processing and validation steps which does not require any patient data sharing. The finetuning is based on the other center's data for further training of the initial model. After finetuning the model, we obtain an average AUC improvement of 13% and 7% with respect to the initial models. This research demonstrates that the predicting capabilities of deep learning models can be extended to and cross-validated with other centers, independent of limitations in data-sharing policies. Moreover, the study shows that finetuning can be exploited to considerably improve the accuracy of the prediction models.

Introduction

Aortic valve stenosis is the most common valvular heart disease in the developed world, impacting dominantly the elderly population (1). If left untreated, the disease has a devastating course, rapidly causing death when symptoms develop. Aortic valve stenosis is commonly caused by calcification of the aortic valve. The treatment for severe aortic valve disease traditionally consists of surgical aortic valve replacement (SAVR). However, in the past decades, transcatheter aortic valve implantation (TAVI) has been developed and approved for use in low- to severe-risk aortic valve disease (2), which is currently a routine treatment worldwide. In fact, recently two randomized controlled trials have shown that the use of TAVI in low-risk patients was non-inferior compared to SAVR (1).

Although TAVI is in continuous development, there are still risks. The broad use of this procedure in the last years has shown a high chance of successful outcomes because of a strict patient selection, which is performed by a multi-disciplinary team. The selection is achieved in an inter-disciplinary discussion meeting and by using planning and treatment support tools, which have recently become available (3–5).

Candidates for TAVI are often frail patients with several comorbidities and with a complicated medical history (1). Despite this, certain patients sometimes do not benefit or only gain limited advantage (6), while it can yield complications during or after the procedure.

Careful patient selection is of paramount importance. Identifying patients who are likely to have improvements, or who are at a higher risk after TAVI, is essential. Improved selection of these patients could reduce mortality after the procedure and would further improve the decision making (1). Moreover, this would lead to an improvement in the treatment efficiency of medical centers and hospitals.

Current risk models have only limited accuracy in predicting TAVI outcomes (7). Multiple models intended for SAVR, thus non-specific TAVI, have been developed and are currently used for risk estimation. Among these are the EuroSCORE, EuroSCORE II and the STS (Society of Thoracic Surgery) scores (8–11). These are procedural or 30-day mortality predictors. A more specific 30-

day mortality predictor is the TAVI-specific TVT registry score (12). One-year mortality has been identified by the medical experts as the life expectancy threshold, above which the TAVI procedure is enabled and appropriate to be performed. However, one-year mortality is even more challenging (13) and current predictors (14)-(15) do not offer optimal results.

While previously mentioned clinical prediction models rely on traditional statistical regression approaches (16), machine learning approaches (ML) have shown competitive predictive value (17) and are capable of benefit from non-linear relationships. Especially when applied to a large amount of data, ML techniques outperform conventional regression models (18). Recently, two ML approaches have been developed and have shown promising results by applying gradient boosting on a decision tree algorithm (GBDT) [18] [19], to predict one-year mortality for patients treated with TAVI. The GBDT techniques were validated on retrospective patient data from single centers without external validation. Moreover, a unique model, exploiting both center data, could provide even more accurate results.

External validations are needed to assess the generalization capabilities of a model on other populations from different centers. This can be performed with GBDT techniques. However, the creation of a unique GBDT-based model, incorporating information data from multiple centers, is technically not possible without merging data from all centers. Exchanging the multi-center data involves specific ethical committees' protocols. This involves long administrative procedures that can sometimes be difficult to achieve, especially in those scenarios where data sharing policies are intrinsically limited (e.g. General Data Protection Regulation).

Deep Learning techniques, which are based on Neural Networks, offer the possibility of updating the model at later stages, as additional training. This process is known as finetuning. With this approach, it is possible to continue the training process of a model with different data. Consequently, data sharing is not needed, since models can be exchanged to be re-trained (finetuned). We have therefore developed TAVI outcome prediction models for one-year mortality while exploiting the finetuning technique.

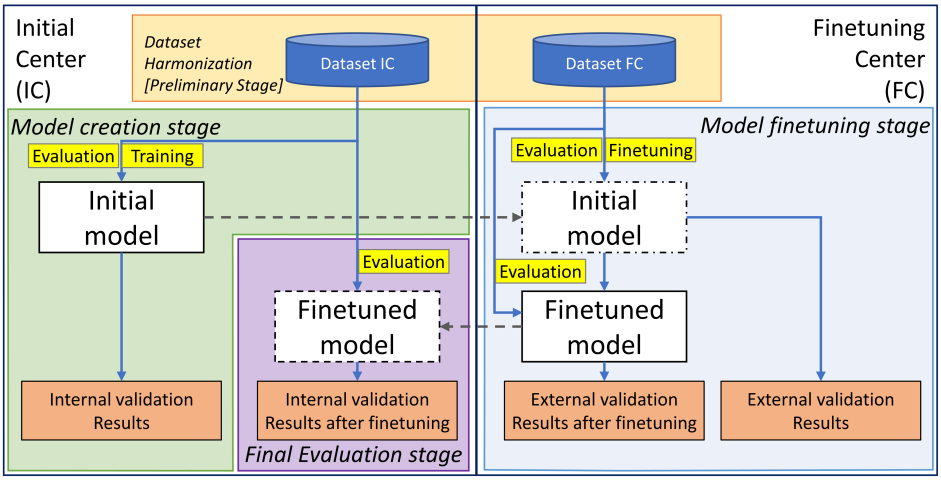

Figure 1. High-level and simplified overview of the cross-validation and finetuning. (Model creation stage) The model is created and internally validated on the IC data and sent to the next stage. (Model finetuning stage) The model is finetuned on the FC data and sent to the next stage. The model is validated on the FC data before and after the finetuning. (Final evaluation stage) The finetuned model is validated on the IC data.

In this paper, we are investigating a prediction model for TAVI where multi-center data are used for cross-validation and leading to a unique model without data sharing while exploiting a finetuning technique.

This work has multiple technical contributions. Firstly, we create two independent mortality prediction models based on DL techniques. Secondly, we cross-validate both models to verify their generalization capabilities, involving two centers with normally different populations. We finetune each model using the new, unseen data of the alternative center for a final validation.

The entire process is achieved by organizing the protocol in three stages, which requires the exchange of the trained and finetuned models for their evaluation, as indicated in the simplified block diagram of Figure 1.

Methods

In this section, the main steps of the inter-center cross-validation protocol are discussed in detail.

High-level overview of the protocol

The inter-center cross-validation protocol is organized in three stages (Figure 2). The first stage is the initial model creation, the second stage is the finetuning of the model, and the third stage contains the final evaluation. Each stage has certain steps and specific inputs and outputs, sometimes shared between both centers, as shown in Table 1. The order of execution is crucial because each output is needed by the successive stage to produce the next output.

Table 1. Input-Output Stages and steps

Stage	Input	Output	Steps
Model creation [1/3]	Dataset IC	Pre-Processing IC Initial Model, Internal validation results	Imputation, Pre-Processing, Oversampling, Training, Evaluation
Model finetuning [2/3]	Dataset FC, Pre-Processing IC, Initial Model	Finetuned Model, External validation results, External validation results after finetuning	Imputation, Pre-Processing, Oversampling, Finetuning, Evaluation
Final Evaluation [3/3]	Imputed and pre-processed Dataset IC, Finetuned Model	Internal validation results after finetuning	Evaluation

The initial model creation stage (Figure 2 left) and the model finetuning stage (Figure 2 right) share a common processing chain including the following steps: imputation, pre-processing, oversampling and training/finetuning, as discussed in detail in Sections 0, 0, 0 and 0, respectively. The final evaluation stage (Figure 2 bottom) is necessary to compare the model before and after the finetuning. At each stage, all created models are evaluated. Section 0 discusses in detail the evaluation step.

Figure 2 details the construction of the protocol diagram, showing all execution steps within each stage. Training, evaluation and finetuning steps are shown in yellow color, while data harmonization, imputation, pre-processing, and oversampling steps are shown in grey color.

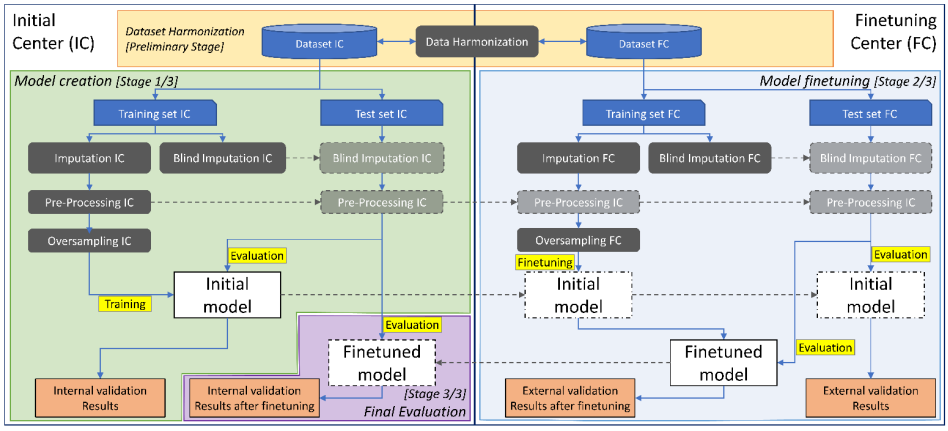

Figure 2. Detailed high-level diagram of the inter-center cross-validation protocol.

Data harmonization

Data harmonization is an important, preliminary and extensive step to align the dataset to a common representation. In this study, we only included patient data that was common (cross-available) to both centers. If the information was represented in a different form, we manually matched the different data to a common representation. Clinical data can be both numerical and categorical. Whereas numerical data can be represented in different units, categorical data is expressed by a different number of instances. In case the number of categories was different, we reduced the number of categories to the amount common to both datasets. The dataset harmonization stage is schematically depicted in Figure 2 (top).

Imputation

As a natural consequence of each data collection process, most datasets contain a certain amount of missing values. To deal with this problem, we discarded all features that had a percentage of missing values higher than 70%. The included features are shown in Table 2. We imputed the missing values using MissForest (21). This is achieved by creating multiple models based on Random Forest to generate an iterative estimation of the missing values, by using non-missing value information until the convergence.

Pre-processing

A step of one-hot encoding is used for categorical features. Numerical features are standardized by removing their mean and by scaling them to unit variance. The mean and the standard deviation values are computed from the training set of the initial center by a scaling function, which is used to standardize the datasets of both centers. After the pre-processing, the clinical data shown in Table 2 resulted in 22 features.

Oversampling

Given the high chance of success of the TAVI procedure, only a small fraction of the patients do not survive the first year. This results in an imbalanced class problem. One common approach used to address imbalanced datasets consists of a random oversampling of the minority class. Here, we have adopted both the random oversampling and the synthetic minority oversampling technique for nominal and continuous (SMOTE-NC) (22), which is based on data interpolation with the *k* nearest-neighbors technique. This approach is tested for the minority class solely and for both classes, thereby augmenting the training set by a factor of three. This strategy has been applied only to the training set, to virtually increase the amount of training data by generating new samples (synthetic patients). This has shown to be useful to increase the accuracy (23) by eventually improving the neural network capabilities of learning new patterns already created by the patient data interpolations and present in the training set.

Training/finetuning

The applied neural network architecture is illustrated in Figure 3. The finetuning of the neural network affects the weights of the connections of both the first and the second dense layer. We used Adam optimizer with a 0.0001 learning rate and 150 epochs for the model creation stage and 100 for the model finetuning stage. The chosen hyper-parameters and architecture were manually selected and a grid search for further optimization was not performed.

Table 2. Summarized patient characteristics from Academic Medical Center (AMC) and Catharina Hospital of Eindhoven (CZE-TU/E).

		AMC		CZE-TU/e	
Clinical name	**Unit or instances**	**Survived (1170)**	**Non-survived (130)**	**Survived (450)**	**Non-survived (66)**
Chronic obstructive pulmonary disease	No	755	66	372	47
	Yes	282	56	76	19
Diabetes	No	719	79	341	46
	Yes	313	43	108	20
Body Mass Index	kg/m^2	28±5 (1134)	27±6 (130)	27±4 (449)	26±5 (66)
Creatinine	µmol/L	98±40 (1126)	119±56 (130)	109±67 (444)	116±46 (66)
Smoking	No	489	59	322	51
	Former	479	49	30	4
	Yes	104	14	98	11
Beta Blockers class of medicine	No	480	42	384	52
	Yes	651	87	66	14
Hemoglobin	mmol/L	7.8±1.0 (1123)	7.7±1.3 (129)	7.9±1.0 (321)	7.5±0.9 (45)
QRS complex time	msec	104±26 (344)	105±31 (52)	110±29 (449)	121±33 (62)
Aortic Valve Area	cm^2	0.8±0.2 (949)	0.8±0.2 (111)	0.7±0.2 (148)	0.7±0.2 (18)
Aortic Valve Peak Gradient	mmHg	68±22 (956)	63±27 (117)	77±24 (172)	71±32 (19)
Aortic Valve mean Gradient	mmHg	68±23 (622)	43±19 (69)	46±16 (137)	39±20 (15)
Previous Myocardial Infarction	No	851	91	297	36
	Yes	187	31	81	15
Sex	Male	514	65	240	45
	Female	656	65	210	21
Age	years	81±7 (1170)	82±10 (130)	80±6 (450)	80±6 (66)
New York Heart Association (NYHA) Functional Classification	1	32	2	3	4
	2	236	20	50	5
	3	506	86	125	23
	4	80	422	33	9
Previous Devices (pacemaker, etc)	No	970	104	312	35
	Yes	102	18	43	12

Evaluation

Two independent experiments have been performed, one considering AMC as the initial center and CZE-TU/e as finetuning center, and one considering CZE-TU/e as the initial center and CZE-TU/e as finetuning center. These paired experiments were necessary to perform the cross-validation and the finetuning with a bi-directional approach to both centers.

The evaluation consisted of twenty-fold cross-validation for both model creation, model finetuning and final evaluation stages. At the model creation stage, this leads to the exchange of 20 scaling functions (Pre-processing IC step in Figure 2) and 20 neural network models (Initial model) per center. At the finetuning stage, this leads to the exchange of 20 finetuned neural network models (Finetuned model) per center. In synthesis, the exchange protocol shown in Figure 2 is performed 20 times, one time per each individual fold.

The training set consists of 95% of the data (of which approximately 10% was used as validation set) and 5% was used as test set, per fold. The validation set was used to analyze the convergence of the neural network by visually identifying possible overfitting of the network and consequently tune its parameters. The validation set is not shown in Figure 2 for simplicity.

All scaling (pre-processing IC) and neural networks models (Initial model) were exchanged between the two centers at the first model creation stage. The finetuned models (Finetuned model) were exchanged once again for the final evaluation stage performed on the initial test set, as shown in Figure 2. A synchronized cloud storage directory was used to facilitate the exchanges of all models and the corresponding scalers. As a result, the validation was performed four times during the three stages, two internal validation and two external validation, both before and after the finetuning, as shown in Table 1. A total of 20 AUCs have been computed per evaluation since twenty folds were used.

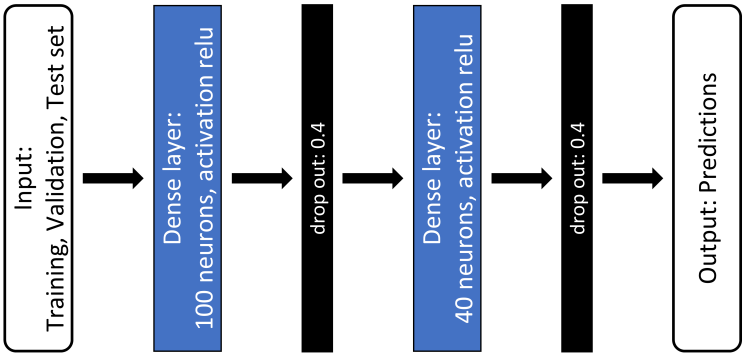

Figure 2. Neural network architecture.

Results

General results overview

In Figure 4 and Figure 5, the results of both the internal validations and external validations are shown. They are represented in the form of a boxplot containing the AUCs of the ROC (AUROC) of the twenty folds for each evaluation. Both figures represent the results before and after the finetuning with respect to the diagram shown in Figure 2.

Figure 4 shows the results of CZE-TU/e executing the model creation and final evaluation stages, and the AMC results of executing the model finetuning stage. Alternatively, in Figure 5 the reciprocal process is shown, thus with AMC executing the model creation and final evaluation stages and CZE-TU/e executing the model finetuning stage.

The results are plotted varying numbers of oversampling as shown on the *x*-axis, where 0 is defined as random oversampling, 1 as SMOTE-NC oversampling applied to the minority class, and 3 as SMOTE-NC augmentation by a factor of three. These results are also shown also in Table 3.

Cross-validation results

The internal validation is the first assessment of the performance of the predictive model on the initial dataset. This is the starting point of the validation process and any other result has to be compared with respect to this starting point. The internal validation results show an average AUC of 0.584 for AMC and 0.544 for CZE-TU/e.

The external validation results show that the models are capable to generalize on the data of another center, with similar accuracy. In fact, an average AUC of 0.588 and 0.583 was observed for AMC and CZE-TU/e, respectively, showing similar results to the internal validation.

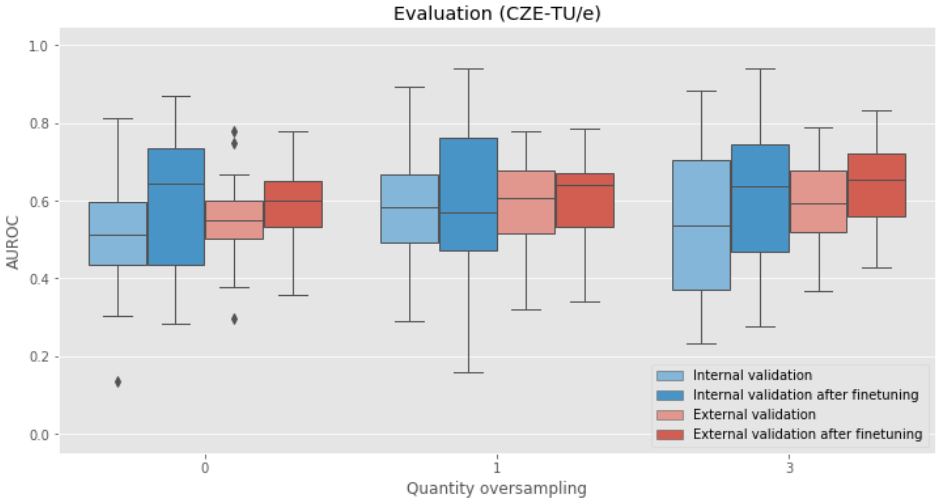

Figure 2. Results for CZE-TU/e as initial center and AMC as finetuning center.

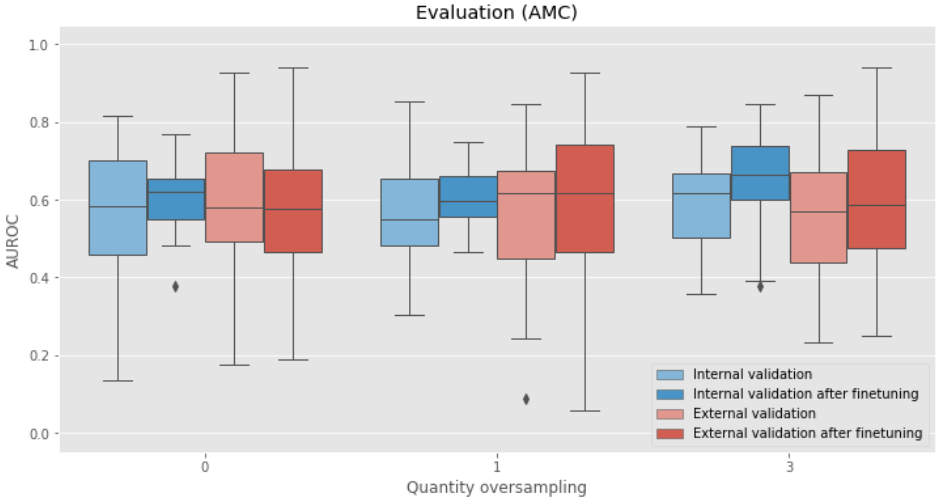

Figure 3. Results for AMC as initial center and CZE-TU/e as finetuning center.

Table 3. Results and mean improvement per evaluation

		Oversampling				
Initial center	**Evaluation**	**0**	**1**	**3**	**Mean**	**Mean Improvement**
AMC	Internal valid.	0.585	0.550	0.619	0.584	+7.30%
	Internal valid. after finetuning	0.619	0.598	0.664	0.627	
	External valid.	0.580	0.617	0.569	0.588	+0.93%
	External valid. after finetuning	0.576	0.619	0.587	0.594	
CZE-TU/e	Internal valid.	0.512	0.584	0.537	0.544	+13.54%
	Internal valid. after finetuning	0.646	0.571	0.638	0.618	
	External valid.	0.548	0.608	0.593	0.583	+8.44%
	External valid. after finetuning	0.601	0.640	0.655	0.632	

Cross-validation results after finetuning

The finetuning increased the accuracy of both the internal validation and external validation. The improvement due to the finetuning is higher for the first experiment where CZE-TU/e is the initial center and AMC is the finetuning center. The results have shown that there is an improvement of 13% on the internal validation and 8% on the external validation.

Similar results are obtained for the second experiment, where AMC is the initial center and CZE-TU/e is the finetuning center. The results have shown that there is an improvement of 7% on the internal validation and 0.9% on the external validation.

Discussion

The cross-validation results have shown similar accuracy for the internal validation and the external validation, for both experiments. This evaluation was needed to assess the predicting potential and generalization capabilities of the one-year mortality model prior to the finetuning.

The comparison of the model accuracies after the finetuning has shown that the finetuning improves the overall accuracy. The models created with the CZE-TU/e

data have shown higher accuracy after the finetuning with respect to the model created with the AMC data. This improvement can be explained by the fact that the AMC data used to finetune the model is based on a larger population than the one of CZE-TU/e, thereby including more variation.

These are important achievements because they motivate further centers to use their clinical data, in the future, to create updated and optimized versions of this or other models possibly yielding higher validation accuracies and gradually more extended validations. To this end, a proper data harmonization is required to align the data between multiple centers.

Adding further clinical data could possibly add more informative features and, theoretically, could lead to better results. Besides this, the architecture of the neural network may be improved and an optimized search for the hyperparameters would be appropriate.

Conclusions

We have developed a dedicated exchange protocol to overcome data exchange and policy-related issues. The proposed protocol enables the cross-validation of two deep learning models used to predict one-year mortality for TAVI. Finetuning was successfully used to improve the results by retraining the model on the dataset of the other cooperating center

This study has shown that finetuning is a promising technique to improve prediction models for the use in new centers and organize this in a cooperative fashion. Moreover, this study provides an exchange protocol, which can be used for other clinical applications and further validation when multiple centers are involved.

References

1. Baumgartner H, Falk V, Bax JJ, De Bonis M, Hamm C, Holm PJ, et al. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. Eur Heart J. 2017 Sep;38(36):2739–91.
2. FDA expands indication for several transcatheter heart valves to patients at low risk for death or major complications associated with open-heart surgery.
3. Khalil A, Faisal A, Lai KW, Ng SC, Liew YM. 2D to 3D fusion of echocardiography and cardiac CT for TAVR and TAVI image guidance. Med Biol Eng Comput. 2017;55(8):1317–26.
4. Grbic S, Mansi T, Ionasec R, Voigt I, Houle H, John M, et al. Image-based computational models for TAVI planning: From CT images to implant deployment. Med Image Comput Comput Interv. 2013;395–402.
5. Swee JKY, Grbić S. Advanced transcatheter aortic valve implantation (TAVI) planning from CT with ShapeForest. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2014. p. 17–24.
6. Puri R, Iung B, Cohen DJ, Rodés-Cabau J. TAVI or No TAVI: Identifying patients unlikely to benefit from transcatheter aortic valve implantation. Eur Heart J. 2016;37(28):2217–25.
7. Martin GP, Sperrin M, Ludman PF, de Belder MA, Gale CP, Toff WD, et al. Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation. Am Heart J. 2017;184:97–105.
8. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg. 2009;88(1):S23–42.
9. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R, et al. European system for cardiac operative risk evaluation (Euro SCORE). Eur J cardio-thoracic Surg. 1999;16(1):9–13.
10. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. Euroscore ii. Eur J cardio-thoracic Surg. 2012;41(4):734–45.
11. The Society of Thoracic Surgeons (STS). STS Web Risk Calculator v2.9.
12. Pilgrim T, Franzone A, Stortecky S, Nietlispach F, Haynes AG, Tueller D, et al. Predicting mortality after transcatheter aortic valve replacement: external validation of the transcatheter valve therapy registry model. Circ Cardiovasc Interv. 2017;10(11):e005481.
13. Ludman PF, Moat N, de Belder MA, Blackman DJ, Duncan A, Banya W, et al. Transcatheter aortic valve implantation in the United Kingdom: temporal trends, predictors of outcome, and 6-year follow-up: a report from the UK Transcatheter Aortic Valve Implantation (TAVI) Registry, 2007 to 2012. Circulation. 2015;131(13):1181–90.
14. Debonnaire P, Fusini L, Wolterbeek R, Kamperidis V, Van Rosendael P, Van Der Kley F, et al. Value of the "TAVI2-SCORe" versus surgical risk scores for prediction of one year mortality in 511 patients who underwent transcatheter aortic valve implantation. Am J Cardiol. 2015;115(2):234–42.
15. Arnold S V, Afilalo J, Spertus JA, Tang Y, Baron SJ, Jones PG, et al. Prediction of poor outcome after transcatheter aortic valve replacement. J Am Coll Cardiol. 2016;68(17):1868–77.

16. Van Mourik MS, Vendrik J, Abdelghani M, Van Kesteren F, Henriques JPS, Driessen AHG, et al. Guideline-defined futility or patient-reported outcomes to assess treatment success after TAVI: What to use? Results from a prospective cohort study with long-term follow-up. Open Hear. 2018;5(2).
17. Jie MA, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;
18. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PDR, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108(11):1723.
19. Mamprin M, Zelis JM, Tonino PA ~L., Zinger S, de With PH ~N. Gradient Boosting on Decision Trees for Mortality Prediction in Transcatheter Aortic Valve Implantation. arXiv e-prints. 2020 Jan;arXiv:2001.02431.
20. Lopes RR, van Mourik MS, Schaft E V., Ramos LA, Baan J, Vendrik J, et al. Value of machine learning in predicting TAVI outcomes. Netherlands Hear J [Internet]. 2019 Sep 20;27(9):443–50. Available from: http://link.springer.com/10.1007/s12471-019-1285-7
21. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.
22. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
23. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput. 2013;3(2):224.

4

Local and distributed machine learning for inter-hospital data utilization: an application for TAVI outcome prediction

Lopes RR*, Mamprin M*, Zelis JM, Tonino PA, van Mourik MS, Vis MM, Zinger S, de Mol BA, Marquering HA.

Frontiers in cardiovascular medicine. 2021:1559.

* Shared first author

DOI: 10.3389/fcvm.2021.787246

Abstract

Machine learning models have been developed for numerous medical prognostic purposes. These models are commonly developed using data from single centers or regional registries. Including data from multiple centers improves robustness and accuracy of prognostic models. However, data sharing between multiple centers is complex, mainly because of regulations and patient privacy issues.

We aim to overcome data sharing impediments by using distributed ML and local learning followed by model integration. We applied these techniques to develop 1-year TAVI mortality estimation models with data from two centers without sharing any data.

A distributed ML technique and local learning followed by model integration was used to develop models to predict 1-year mortality after TAVI. We included two populations with 1,160 (Center A) and 631 (Center B) patients. Five traditional ML algorithms were implemented. The results were compared to models created individually on each center.

The combined learning techniques outperformed the mono-center models. For center A, the combined local XGBoost achieved an AUC of 0.67 (compared to a mono-center AUC of 0.65) and, for center B, a distributed neural network achieved an AUC of 0.68 (compared to a mono-center AUC of 0.64).

This study shows that distributed ML and combined local models techniques, can overcome data sharing limitations and result in more accurate models for TAVI mortality estimation. We have shown improved prognostic accuracy for both centers and can also be used as an alternative to overcome the problem of limited amounts of data when creating prognostic models.

Introduction

Transcatheter Aortic Valve Implantation (TAVI) is a consolidated procedure for aortic stenosis treatment. To support patient selection, traditional risk stratification models, either for general cardiac surgery or TAVI specific, are used for mortality estimation (1,2). Other models, exploiting more complex algorithms, have shown higher accuracies when compared to traditional logistic regression-based models (3,4). Nevertheless, mortality estimation models have shown limited prediction accuracy when tested on other center's populations than the one used to generate the models (5–8). This can be explained by the different distribution in the populations, given by different patient selection or practice variation among institutions.

Mitigating the models' accuracy drop on different populations is essential to obtain models with higher generalization capability. For this purpose, model updating or fine-tuning have been used successfully (9,10). These techniques consist of making small adjustments in the model, using data from a different population, to make the models more robust for that specific population and achieve higher accuracies. It is also known that machine learning (ML) models usually benefit from a large amount of data, allowing to learn complex non-linear interactions among variables. Ideally, a single model would be developed using data from multiple centers to optimize the model's accuracy. As a practical alternative, models can be iterated by making small adjustments for each population. Sharing data between centers, however, is a complex procedure because of regulations dealing with patient's privacy and, therefore, this is not always possible in practice because of data protection regulations such as the European General Data Protection Regulation (11).

One possible approach to overcome the data sharing limitation is by exploiting distributed ML techniques. These techniques allow the training of models at multiple physical locations, regardless of their geographical distance, with limited or no data sharing. A popular distributed ML strategy, called Cyclical Weight Transfer (CWT), consists of sharing a single model across locations sequentially and cyclically for incremental updates. At each location, the model is modified using the data available at that center before sending it to the next location. This approach has been used to train deep learning models with medical

images, achieving similar results as if the data was located in a single location (12). A simpler approach is to combine models trained locally at different locations. This can be achieved by using stacking ensemble, where the prediction probabilities of the models trained locally are used as features to fit a logistic regression (LR) model (13,14). With these approaches, the models are expected to have a higher reliability and achieve better generalization capability.

In this study, we exploited two techniques to deal with the data sharing limitation to potentially improve the accuracy of models for 1-year TAVI mortality prediction. To this end, we trained multiple models based on CWT and stacking approaches across two centers without data sharing.

Methods

Population

Models to predict 1-year modality were created with data from a total of 1,791 patients who underwent TAVI procedures in two distinct centers were included in this study. The Amsterdam UMC—Location AMC (AMC) with 1,160 consecutive patients (first dated October 2007 and last dated April 2018) and the Catharina Hospital of Eindhoven (CZE) with 631 consecutive patients (first dated January 2015 and last dated December 2018). The 1-year mortality information was collected from a follow-up study for the AMC and by the national census for the CZE. Patients with missing outcome or with more than 50% of missing data were excluded from the study. This study, considering also where the data were located, was performed at the Amsterdam UMC and the Eindhoven University of Technology for the CZE.

Pre-processing

Only variables that were available in both datasets were included while missing values were imputed with the mean for numerical variables and the mode for categorical variables. The measures of central tendency used for imputation were calculated for each center and used to impute it owns center's data.

Additional pre-processing was applied to the data for the development of the Neural Networks (NN) to facilitate its convergence. The continuous variables were standardized by removing their mean and by scaling them to unit variance

while one-hot encoding was applied for categorical features. These steps are not required for the other classifiers.

Two approaches were evaluated to deal with the class imbalance during training: class weighting and random over-sampling. The first approach consists of assigning different weights to balance the loss of the two classes during training. The second approach consists of randomly oversampling the samples of the minority class.

Model Development

In this study, we evaluated four distinct classifiers: Random Forest (RF), Extreme Gradient Boosting (XGB) (15), CatBoost (CATB) (16), and NN. Two NN architectures were evaluated: a narrow and a wide. The narrow is composed by two layers of 8 and 4 neurons while the wide is composed by two layers with 100 and 40 neurons. The complete architectures are described in the Supplementary Material Table I. All experiments were performed on Python 3.6.9 and scikit-learn library 0.21.3 (17). We used a CWT approach to train the models with data from both centers in an iterative fashion. Besides that, we also evaluated stacking models trained individually for each center. For this, prediction outputs from models trained on each center were used to train a LR model and obtain a unique prediction output.

Cyclical Weight Transfer Approach

The CWT approach is slightly different for the NN and the tree-based models. In CWT, as illustrated in Figure 1, the NN weights are initialized by one center and sent to the other center for updating the weights with the other center's data. This updating procedure continues until the stopping criteria is reached. Dropout was included between layers to randomly prevent some neurons from being updated by the training center (18).

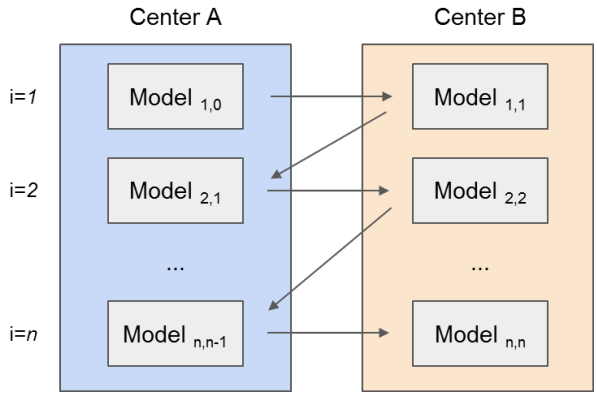

Figure 1. Illustration of the development of a prognostic model using the Cyclic Weight Transfer approach. The model is being trained by two different centers in an iterative fashion. Each model version is trained and exchanged between centers for *n* iterations or until a stopping criterion is reached.

The tree-based models (RF, CATB, and XGB) were trained by adding new trees, from each center, at every new iteration. To this end, the models were exchanged iteratively between centers resulting in the forest to grow. For example, as illustrated in Figure 2, an initial model created with a single tree for the first center is sent to the second center, where a new tree is added. This exchanging iterative process continues until the stopping criterion is reached: a maximum number of iterations (500) or the validation error stopped decreasing for both centers after 10 epochs. Although the trees created by one center are never modified by the other, the model is iteratively being updated by the addition of new trees from each center. For the XGB and CATB, the previous trained trees are taken in consideration when fitting new trees.

The center with the largest amount of data was used to start the training process. The hyperparameters and architectures were empirically optimized. Information regarding the values for which the hyperparameter optimization was performed can be found in the Supplementary Material Table II.

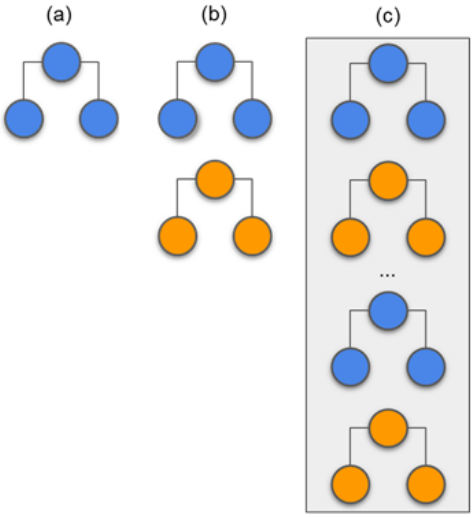

Figure 2. Example of a tree-based model (random forest) that is created by two different centers (represented by different colors). (a) A predefined number of trees is initially created by the first center. (b) The second center adds new trees to the forest, without modifying the previous trees. (c) The random forest training process is complete, with the same number of trees from each center.

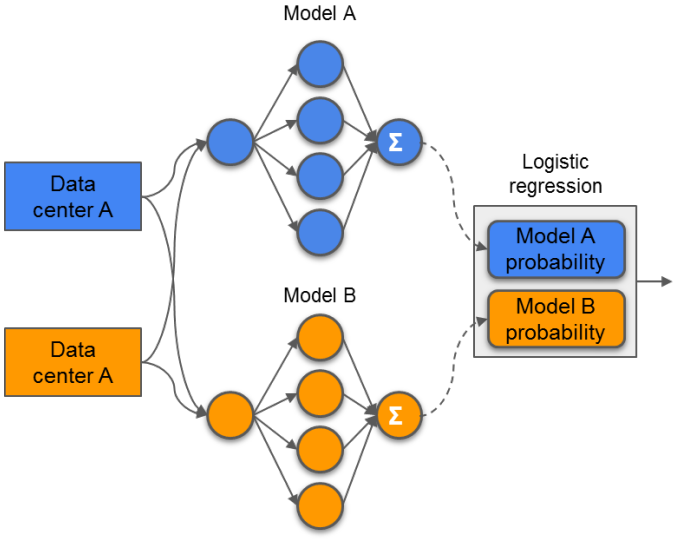

Figure 3. Example of a stacking model. The models are trained independently on each center and its prediction probabilities are used as features to train a single logistic regression model.

Stacking Approach

Stacking has been successfully applied in previous studies (19,20). At the initiation of the process, the models were trained locally at each center. To this, the hyperparameters were optimized via grid search with 5-fold cross-validation. The evaluated hyperparameters are presented in the Supplementary Material Table III. After both centers had their models trained, they were used to compute the probability output for all samples (training and testing). The probability output from both center's training set was used as features (2 features in total; the probability from center A and the probability from center B) to train an LR model. The probability output from the test samples was used to evaluate the LR model. With this approach, represented in Figure 3, the models and probability outputs from both centers were exchanged only once. Different classifiers were not stacked together (i.e., the NN from center A was only combined with the NN from center B).

Internal Evaluation

To evaluate the value of creating models using data from 2 centers, we compared these models with the models that were trained on the data from only 1 center. These mono-center models were trained locally and tested on its own data. The optimization and evaluation of these models was the same as the used for the stacked approach, with hyperparameter optimization via grid search and evaluation with a 5-fold cross-validation scheme. These models have already been developed in a previous study (7).

Evaluation

The models were evaluated with stratified 20-fold cross-validation. With this, each center split its own data in 20-folds, leading to twenty iterations with different test sets. The testing folds were kept unused until the final evaluation. The area under the curve (AUC) of the receiver operating characteristic (ROC) was used to evaluate each model. The average of the twenty AUCs, as well as the standard deviation (std), was reported for each center.

Results

Among all 1,791 patients from two centers include on this study, 188 patients (10%) did not survive through the first year after TAVI. The baseline characteristics of the patients from both centers are summarized in Table 1.

The cyclical NN model with a narrow architecture achieved the highest average ROC AUC of 0.66 (center A: 0.64 AUC, center B: 0.68 AUC). This NN also achieved the highest score for center A. The stacked models with the highest accuracies achieved a ROC AUC of 0.65. This accuracy was achieved by three models; CATB (center A: 0.64 AUC, center B: 0.65 AUC), XGB (center A: 0.67 AUC, center B: 0.63 AUC) and the NN with a narrow architecture (center A: 0.64 AUC, center B: 0.65 AUC). The stacked XGBoost achieved the highest individual accuracy for center B. In Figure 4 we show the average ROC of the models with highest AUCs and in Table 2 we present all results.

The highest average accuracy for the mono-center models was a ROC AUC of 0.64, achieved by CATB, RF and the NN with narrow architecture. The highest individual accuracy was achieved by XGB for center A (AUC of 0.65) and CATB for center B (AUC of 0.64).

Discussion

Our proposed approaches of distributed and combined local models to predict 1-year TAVI mortality with data from two centers outperformed the models trained with each center individually (mono-center). T approaches do not require the data to be sent from center to center once each center process its own data. Additionally, the centers benefited from training the models using these approaches, once their accuracies outperformed the accuracies of the mono-centers models (trained locally and independently). For both centers, the combined prediction models outperformed the models using only the local data. These approaches can be extended to multiple centers or different problems, not being exclusive for TAVI.

Table 1. Descriptive statistics of the study group, mean ± SD or N (%).

		Center A		Center B	
Variable	**Instances**	**Survived (n=1039)**	**Non-survived (n=121)**	**Survived (n=564)**	**Non-survived (n=67)**
Sex	*Male*	587 (56%)	61 (50%)	303 (54%)	46 (69%)
	Female	452 (44%)	60 (50%)	261 (46%)	21 (31%)
Age (year)		81 ± 7	83 ± 7	81 ± 6	80 ± 6
Chronic obstructive pulmonary disease	*No*	755 (73%)	66 (55%)	464 (83%)	48 (71%)
	Yes	282 (27%)	55 (45%)	98 (17%)	19 (28%)
Diabetes	*No*	720 (69%)	79 (65%)	241 (75%)	47 (70%)
	Yes	313 (30%)	42 (35%)	142 (25%)	20 (30%)
Body mass index (kg/m2)		28 ± 5	27 ± 6	27 ± 4	26 ± 4
Creatinine (µmol/L)		98 ± 41	120 ± 56	108 ± 62	116 ± 46
Smoking	*No*	479 (46%)	59 (49%)	392 (70%)	51 (76%)
	Former	456 (44%)	49 (40%)	41 (7%)	4 (6%)
	Yes	104 (10%)	13 (11%)	131 (23%)	12 (18%)
Beta blockers class of medicine	*No*	596 (57%)	80 (66%)	498 (88%)	53 (79%)
	Yes	437 (42%)	40 (33%)	66 (12%)	14 (21%)
Hemoglobin (mmol/L)		7.8 ± 1	7.7 ± 1	7.9 ± 1.0	7.5 ± 0.9
QRS complex time (msec)		104 ± 26	107 ± 27	110 ± 29	121 ± 33
Aortic valve area (cm2)		0.8 ± 0.2	0.8 ± 2	0.7 ± 0.2	0.7 ± 0.2
Aortic valve peak gradient (mmHg)		68 ± 23	64 ± 26	77 ± 24	71 ± 32
Aortic valve mean gradient (mmHg)		43 ± 16	44 ± 19	46 ± 16	39 ± 20
Previous myocardial infarction	*No*	851 (82%)	91 (75%)	387 (79%)	36 (69%)
	Yes	187 (18%)	30 (25%)	105 (21%)	16 (31%)
New York Heart Association (NYHA) functional classification	*1*	28 (3%)	1 (1%)	9 (3%)	4 (9%)
	2	220 (21%)	17 (14%)	59 (20%)	5 (12%)
	3	473 (46%)	82 (68%)	184 (61%)	24 (57%)
	4	76 (7%)	21 (17%)	49 (16%)	9 (21%)
Previous devices (such as pacemaker)	*No*	937 (90%)	104 (86%)	417 (89%)	36 (75%)
	Yes	102 (10%)	17 (14%)	52 (11%)	12 (25%)

Table 2. Average area under the receiver operating characteristic curve and its standard deviation for all experiments. The rows are the classifiers on different setups (cyclical, stacking or internal validation) and the columns are different balancing techniques per center. Highest accuracies per center and on average are highlighted in bold.

		Center A (n=1160)		Center B (n=631)		Average of centers	
	Model	Balanced class weight	Random oversam pling	Balanced class weight	Random oversam pling	Balanced class weight	Random oversam pling
Cyclical	XGB	0.58 ± 0.10	0.58 ± 0.10	0.62 ± 0.16	0.54 ± 0.15	0.60 ± 0.13	0.56 ± 0.13
	CATB	0.62 ± 0.15	0.60 ± 0.14	0.61 ± 0.14	0.61 ± 0.16	0.62 ± 0.15	0.61 ± 0.15
	RF	0.62 ± 0.11	0.61 ± 0.12	0.64 ± 0.13	0.64 ± 0.14	0.63 ± 0.12	0.63 ± 0.13
	NN wide	0.62 ± 0.14	0.63 ± 0.14	0.67 ± 0.14	0.65 ± 0.17	0.65 ± 0.14	0.64 ± 0.16
	NN narrow	**0.64 ± 0.12**	0.62 ± 0.13	**0.68 ± 0.12**	0.62 ± 0.15	**0.66 ± 0.12**	0.62 ± 0.14
Stacking	XGB	**0.67 ± 0.10**	0.61 ± 0.08	0.63 ± 0.17	0.60 ± 0.13	**0.65 ± 0.14**	0.61 ± 0.11
	CATB	0.64 ± 0.11	0.62 ± 0.10	0.65 ± 0.16	0.62 ± 0.13	**0.65 ± 0.14**	0.62 ± 0.12
	RF	0.63 ± 0.10	0.60 ± 0.09	0.64 ± 0.15	0.63 ± 0.15	0.64 ± 0.13	0.62 ± 0.12
	NN wide	0.64 ± 0.13	0.62 ± 0.13	0.64 ± 0.14	0.61 ± 0.11	0.64 ± 0.14	0.62 ± 0.12
	NN narrow	0.64 ± 0.12	0.65 ± 0.13	**0.66 ± 0.14**	0.59 ± 0.14	**0.65 ± 0.13**	0.62 ± 0.14
Mono-center	XGB	**0.65 ± 0.11**	0.59 ± 0.11	0.59 ± 0.17	0.56 ± 0.18	0.62 ± 0.14	0.58 ± 0.15
	CATB	0.63 ± 0.11	0.59 ± 0.12	0.60 ± 0.15	**0.64 ± 0.17**	0.62 ± 0.13	0.62 ± 0.15
	RF	0.65 ± 0.10	0.59 ± 0.11	0.62 ± 0.14	0.62 ± 0.16	**0.64 ± 0.12**	0.61 ± 0.14
	NN wide	0.64 ± 0.11	0.62 ± 0.13	0.63 ± 0.15	0.61 ± 0.15	**0.64 ± 0.13**	0.62 ± 0.14
	NN narrow	0.63 ± 0.12	0.58 ± 0.12	0.65 ± 0.16	0.60 ± 0.16	**0.64 ± 0.14**	0.59 ± 0.14

XGB XGBoost, *CATB* CatBoost, *RF* Random Forest, *NN* Neural network

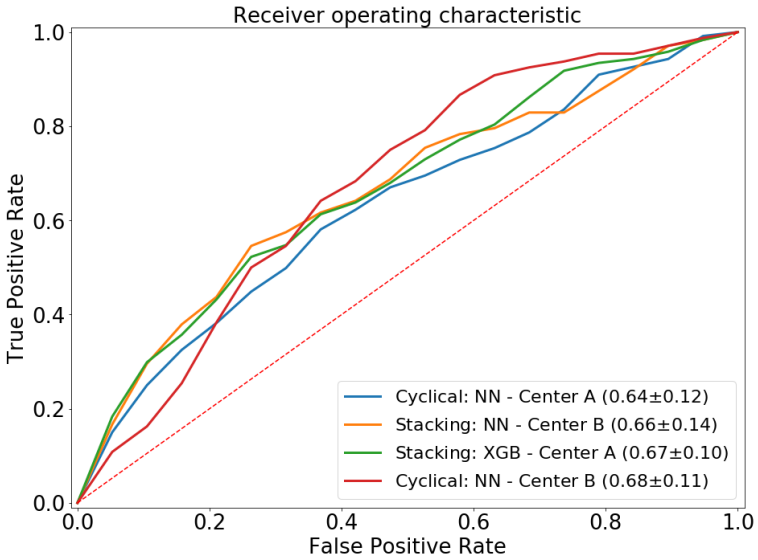

Figure 4. Average ROC curve (standard deviation) of the 20-fold cross-validation for the distributed and combined local models. NN = neural network, XGB = XGBoost.

Some recent studies presented ML models for TAVI outcome prediction. In previous studies, Lopes et al. (3) and Mamprin et al. (4) developed pipelines for outcome prediction for individual centers. Additionally, Al-Farra et al. (6) and Mamprin et al. (7) showed the accuracy drop on the evaluation of previous traditional risk scores or recent ML models when evaluated on different populations. The importance of model updating was highlighted by Lopes et al. (9) and Al-Farra et al. (10), where NN and LR models were updated after the training process was complete. They concluded that model updating is of utmost importance when using the models on different (external) populations.

This study suffered from some limitations. Some important features, which have shown prognostic value in previous studies, were not included in this study because these were not similarly reported by both centers. Also, to be aligned with previous studies, a simple imputation technique was used instead of a multiple imputation. Additionally, although center A has almost twice the number of patients from center B, the data acquisition period is relatively large (11 years, compared to 4 years from center B). This might affect the accuracy of the models since the TAVI procedures are constantly improving, from patient selection to the procedure itself, and the effects of these changes are not included in the

models. Regarding the distributed experiments, the hyperparameter optimization process was reduced to a limited number of options and not many optimizations were implemented since this was not the subject of the study. Numerous additional settings could be adjusted for cyclical training: for example, the NN could be trained for multiple epochs or on mini-batches, weights could be assigned to the loss to deal with different population sizes, or even a combined loss could be taken into account when back-propagating the loss.

Conclusion

In our study, we demonstrate two approaches to overcome the data sharing limitations between medical centers. For both centers, the combined models outperformed models in which only patients from their own center was used: for the larger center, the stacking approach showed the highest accuracy and for the smaller center, the distributed approach achieved the highest accuracy. The highest accuracy improvement was achieved for the center with a smaller number of patients, showing that when limited amounts of data are involved in creating prognostic ML models, federated can be successful option to generate a unique model in a cooperative fashion.

References

1. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. Euroscore ii. Eur J cardio-thoracic Surg. 2012;41(4):734–45.
2. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 2-Isolated Valve Surgery. Ann Thorac Surg [Internet]. 2009;88(1 SUPPL.):S23–42. Available from: http://dx.doi.org/10.1016/j.athoracsur.2009.05.056
3. Lopes RR, van Mourik MS, Schaft E V., Ramos LA, Baan J, Vendrik J, et al. Value of machine learning in predicting TAVI outcomes. Netherlands Hear J [Internet]. 2019 Sep 20;27(9):443–50. Available from: http://link.springer.com/10.1007/s12471-019-1285-7
4. Mamprin M, Zelis JM, Tonino PAL, Zinger S. Decision Trees for Predicting Mortality in Transcatheter Aortic Valve Implantation. Bioengineering. 2021;8(2):22.
5. Martin GP, Sperrin M, Ludman PF, de Belder MA, Gale CP, Toff WD, et al. Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation. Am Heart J. 2017;184:97–105.
6. Al-Farra H, Abu-Hanna A, de Mol BAJM, Ter Burg WJ, Houterman S, Henriques JPS, et al. External validation of existing prediction models of 30-day mortality after Transcatheter Aortic Valve Implantation (TAVI) in the Netherlands Heart Registration. Int J Cardiol. 2020;317:25–32.
7. Mamprin M, Lopes RR, Zelis JM, Tonino PAL, van Mourik MS, Vis MM, et al. Machine Learning for Predicting Mortality in Transcatheter Aortic Valve Implantation: An Inter-Center Cross Validation Study. J Cardiovasc Dev Dis. 2021;8(6):65.
8. Wolff G, Shamekhi J, Al-Kassou B, Tabata N, Parco C, Klein K, et al. Risk modeling in transcatheter aortic valve replacement remains unsolved: an external validation study in 2946 German patients. Clin Res Cardiol. 2021;110(3):368–76.
9. Lopes RR, Mamprin M, Zelis JM, Tonino PAL, van Mourik MS, Vis MM, et al. Inter-Center Cross-Validation and Finetuning without Patient Data Sharing for Predicting Transcatheter Aortic Valve Implantation Outcome. 2020 IEEE 33rd Int Symp Comput Med Syst [Internet]. 2020 Jul;591–6. Available from: https://ieeexplore.ieee.org/document/9183069/
10. Al-Farra H, de Mol BAJM, Ravelli ACJ, Ter Burg W, Houterman S, Henriques JPS, et al. Update and, internal and temporal-validation of the FRANCE-2 and ACC-TAVI early-mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) using data from the Netherlands heart registration (NHR). IJC Hear Vasc. 2021;32:100716.
11. Voigt P, Von dem Bussche A. The eu general data protection regulation (gdpr). A Pract Guid 1st Ed, Cham Springer Int Publ. 2017;10:3152676.
12. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Informatics Assoc. 2018;25(8):945–54.
13. Wolpert DH. Stacked generalization. Neural networks. 1992;5(2):241–59.
14. Tsoumakas G, Vlahavas I. Effective stacking of distributed classifiers. In: Ecai. 2002. p. 340–4.

15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;785–94.
16. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv Prepr arXiv181011363. 2018;
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012;12:2825–30.
18. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
19. Wang Y, Wang D, Geng N, Wang Y, Yin Y, Jin Y. Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. Appl Soft Comput. 2019;77:188–204.
20. Kim C, You SC, Reps JM, Cheong JY, Park RW. Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. J Am Med Informatics Assoc. 2021;28(6):1098–107.

Supplementary Material

Table I. Evaluated neural networks architectures.

Architecture	Layer	Extra param
Narrow	Dense(8)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	LeakyReLU(0.01)	
	Dropout(0.5)	
	Dense(4)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	LeakyReLU(0.01)	
	Dropout(0.5)	
	Dense(1)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	Sigmoid()	
Wide	Dense(100)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	LeakyReLU(0.01)	
	Dropout(0.5)	
	Dense(40)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	LeakyReLU(0.01)	
	Dropout(0.5)	
	Dense(1)	kernel reg.=l2(0.001), activity reg.=l2(0.001)
	Sigmoid()	

Table II. Hyperparameter used and searched to train the distributed models.

Classifier	Param	Value
Tree-based	Trees	1, 3, 5
	Depth	4, 5, 6
Neural networks	Learning rate	0.01, 0.001, 0.0001
	Optimizer	Adam
	Min. epochs	10
	Max. epochs	500
	Early stopping	10

Table III. Hyperparameters grid used for RF, XGB, CATB, and NN.

Classifier	Parameter name	Parameter value
RF	Number of trees	[20, 50, 100, 200]
	Max features	[auto]
	Max depth	[2, 3, 4, 8]
	Min samples per split	[2, 4, 6]
	Min samples per leaf	[1, 2, 4]
	Class weight	[Balanced]
XGB	Number of trees	[500]
	Max features	[None, auto]
	Max depth	[2, 3, 4]
	Gamma	[0, 0.5, 1, 3]
	Subsample	[0.7, 1]
	Learning rate	[0.1, 0.01, 0.001]
	Col sample by tree	[0.7, 1]
	Scale pos weight	[1, 2, 3]
	Min child weight	[1, 5, 10]
CATB	Number of trees	[500]
	Max depth	[2, 3, 4]
	Gamma	[0, 0.5, 1, 3]
	L2 leaf reg	[3, 10]
	Learning rate	[0.05, 0.10, 0.15]
	Auto class weights	[Balanced]
NN	Learning rate	0.01, 0.001, 0.0001
	Architecture	[Narrow, Wide]

5

Temporal validation of 30-day mortality prediction models for transcatheter aortic valve implantation using statistical process control

An observational study in a national population

Lopes RR, Yordanov TTR, Ravelli AACJ, Houterman S, Vis MM, de Mol AJM, Marquering HA, Abu-Hanna A.

Submitted for publication.

Abstract

Various mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) have been developed in the past years. The effect of time on the performance of such models, however, is unclear given the improvements in the procedure and changes in patient selection, potentially jeopardizing the usefulness of the prediction models in clinical practice. We aim to explore how time affects the performance and stability of different types of prediction models of 30-day mortality after TAVI. We developed both parametric (Logistic Regression) and non-parametric (XGBoost) models to predict 30-day mortality after TAVI using data from the Netherlands Heart Registration. The models were trained with data from 2013 to the beginning of 2016 and Statistical Process Control was used to analyse how time affects the models' performance on independent data from the mid of 2016 to the end of 2019. The area under the Receiver Operating Characteristics curve (AUC) was used to evaluate the models in terms of discrimination and the Brier Score (BS), which is related to calibration, in terms of accuracy of the predicted probabilities. To understand the extent to which reupdating the models contribute to the models' stability, we also allowed the models to be updated over time. We included data from 11,291 consecutive TAVI patients from hospitals in the Netherlands. The parametric model without re-training had a median AUC of 0.64 (IQR 0.54-0.73) and BS of 0.028 (IQR 0.021-0.035). For the non-parametric model, the median AUC was 0.63 (IQR 0.48-0.68) and BS was 0.027 (IQR 0.021-0.036). Over time, the developed parametric model was stable in terms of AUC and unstable in terms of BS. The non-parametric model was considered unstable in both AUC and BS. Repeated model updates resulted in stable models in terms of AUC and decreased the variability of BS, although BS was still unstable. The updated parametric model had a median AUC of 0.66 (IQR 0.57-0.73) and BS of 0.027 (IQR 0.020-0.035) while the non-parametric model had a median AUC of 0.66 (IQR 0.57-0.74) and BS of 0.027 (IQR 0.023-0.035). The temporal validation of the TAVI 30-day mortality prediction models showed that the models updated over time are more stable and accurate when compared to the frozen models. This highlights the importance of repeatedly updating models over time to improve or at least maintain their performance stability. The non-parametric approach did not show improvement over the parametric approach.

Introduction

Aortic stenosis is the most common valvular disease in developed countries. If symptomatic, the stenosis requires valve intervention (1). Transcatheter Aortic Valve Implantation (TAVI) has become the routine treatment for aortic stenosis even for low and intermediate risk patients (2–4). Besides the improvements of the procedure and technology involved (5,6), such as using smaller sheaths and organizing specialized teams for the procedure, a strict patient selection is being followed to select patients who are likely to benefit from TAVI (7).

The TAVI candidate selection is performed by a multi-disciplinary team, where multiple risk scores, such as STS (Society of Thoracic Surgery) (8) and EuroSCORE (9,10) are considered. Although those scores are not TAVI specific, they are well accepted parametric models and used for early-mortality estimation after cardiac surgery. Other instruments, such as FRANCE2 (11), ACC-TAVI (12), and also non-parametric models (13,14) aimed at predicting mortality specifically after TAVI have been introduced. The external validation of such models, in which patients originate from other settings and countries, has shown worse performance than on the internal validation obtained on the original dataset (15–17). Such models only achieved improved performance when updated for that specific centre (18,19).

The TAVI patient selection process and the procedure itself are changing over time, and it is still unknown if there is a performance drift in the accuracy of the mortality prediction models over time in the same setting. With that, it is not clear if the prediction of models developed a while ago are stable and fit for continuous use without updates. Although a limited prospective validation was performed in a previous study (18), only a single test set was used by the authors and the performance change and model's stability were not assessed repeatedly over time. In addition, the evaluated risk scores were developed using a parametric model and it is unclear how non-parametric models (such as boosting trees) behave on the same TAVI mortality prediction task. Therefore, an investigation is needed to assess the stability of models over long periods of time. Statistical Process Control (SPC) is a monitoring and alerting instrument that combines graphical and statistical inferences that can be used to monitor the accuracy and errors of the prediction models over time (20,21). With this approach, the model's

stability over time can be visualized and statistically assessed. We aim to explore how time affects the performance and stability of both a parametric and a non-parametric prediction model for 30-day mortality after TAVI. To this end, we used a large dataset from all heart centres in the Netherlands to train the models and use SPC to monitor their stability and performance prospectively.

Methods

Study population

We included all patients registered in the Netherlands Heart Registration (NHR)[1] who underwent a TAVI procedure between January 2013 and December 2019 in the Netherlands. The NHR is a national registry that includes data from all the sixteen-heart intervention centres in the Netherlands, containing demographics, clinical characteristics, intervention, and procedure details (22). The NHR Transcatheter Heart Valve Interventions registration committee gave permission for this analysis in January 2021.

For this study, the outcome used is the 30-day mortality after the TAVI procedure. Two of the sixteen centres were excluded given that less than 5% of the 30-day mortality status of their patients was available when conducting the study. In addition, patients without a mortality status or that had a concomitant procedure (e.g., pacemaker implantation) were not included.

In order to analyse the studied population, general statistics were computed for all variables. Mean and standard deviation was computed to data with normal distribution and median and interquartile range for non-normally distributed data. Chi-square or Two-sample T-test was used as appropriate.

Variables

We included all variables that were available in the NHR and that had been already used in TAVI risk scores and other studies (13–19). Among the variables, demographic data such as age, sex, and body mass index (BMI) were included. Also, clinical history and screening variables, including the estimated Glomerular Filtration Rate (eGFR), the New York Heart Association (NYHA) score, chronic

[1] https://nederlandsehartregistratie.nl

lung disease, dialysis, systolic pulmonary arterial pressure, creatinine, diabetes mellitus (DM), left ventricular ejection fraction, and recent myocardial infarction were included. In terms of the procedure, its acuity, the chosen access route, critical preoperative state and, year of the procedure were included. All used variables were acquired before the TAVI procedure was performed.

The eGFR and creatinine were clipped for values larger than 60 mL/min/1.73m2 and 250 μmol/L, respectively based on expert opinion. Also, DM was represented by three categories: no DM, with untreated DM, and with DM being treated with insulin. Additionally, the procedure acuity and access route were dichotomized to elective/non-elective and femoral/non-femoral access respectively.

To deal with the missing values, an iterative multiple imputation method (MissForest) was used to impute data. For this step, only the data from the training set was used to train the imputation model, which was later used to impute the data on the test set. Dummy variables were created by leaving one category out for the NYHA score, year of the procedure, and DM categories.

Prediction models

We evaluated two well-established parametric and non-parametric techniques: logistic regression (LR) and extreme gradient boosting (XGB). LR is a parametric approach and has one coefficient assigned for each variable of the model, allowing a relatively easy interpretation and low model complexity. On the other hand, XGB is a non-parametric approach, based on building an ensemble of decision trees. With that, predictions of multiple trees are combined into a single prediction. The models were developed using the scikit-learn (23) and XGBoost (24) Python libraries.

Both models had their hyperparameters tuned using a grid-search approach with the training data in a stratified 10-fold cross-validation (CV). Specifically, different sets of parameters were assessed to find the optimum model, such as the error used for training (L1 or L2) of the LR model, and the tree depth for XGB. All hyperparameters assessed are listed in the supplementary material Table S1. The hyperparameter set with the highest average Area Under the Curve (AUC) of the receiver operating characteristics' curve across the tuning data, which is held out of the training set, of all folds was selected and used to train the models.

In order to visualize the agreement between predicted mortality risk and real mortality, a calibration plot was created for all prediction models.

Model validation

Internal validation

Internal validation was performed to evaluate the models regardless of any temporal shifts in the data. To this end, the data from all the treatment years were gathered together and a 10-fold CV was conducted as described above. The imputation model was created based on the training folds and later used to impute the corresponding test set.

The average AUC, with standard deviation (SD), was used to evaluate the models. While the AUC is commonly used for the evaluation of clinical models, it is not sensitive to changes in the prevalence of the event. Hence, the Brier Score (BS), which is sensitive to prevalence and calibration was also selected as a measure of the accuracy of the predicted probabilities. The higher the AUC and the lower the BS, the better.

Temporal validation

Temporal validation was conducted to simulate the models' predictive performance over time, reproducing how they would perform if used in a real-life scenario with prospective patients. To this end, all patients were gathered together and sorted by their procedure date. To have a sufficiently large number of patients and still sizable groups suitable for SPC, the data was split into 38 mutually exclusive groups. Except for the first group, with 320 samples, all remaining groups had 297 samples each.

Group analysis

For context, we first visualize changes in 30-day mortality ratio and age over time. Then, we prepare the data for the SPC analysis. Specifically, from all thirty eight, the first eight groups in which performance was stable were used to train the initial models and to calculate the standard deviation in the performance measures. The standard deviations are used to determine the limits of the statistical control charts. The remaining groups are used for obtaining and scrutinizing the models' performance: based on the SPC limits, one can interpret whether the performance measurements exhibit natural (expected) variability, or

structural (unexpected) variability. Two experiments were performed: a) without re-training the model on each iteration (frozen model) and b) re-training the model on each successive iteration (model update). This is done to compare the stability between the fixed model and a model with a repeated update over time. For the frozen model, the model was trained once and evaluated on all subsequent parts individually. The model update, on the other hand, has the testing data from the previous iteration added to the training data on each new iteration. Figure 1 shows a representation of both experiments.

Statistical process control

SPC is a graphical framework showing the progression of a key measurement over time and, additionally, provides simple rules to judge whether the variation in the measurements reflects expected natural variation or a structural change. In this case, the analysed process is the validation of the parametric and non-parametric TAVI mortality prediction models. A part of the data, 8 groups in this study, as is often recommended, was used to calculate the mean and standard deviation of the process, which are used in judging the nature of the variation. However, performance on these initial groups should show a stable process without trends. Otherwise, subsequent groups, one by one, are used to replace the previous (unstable) groups until performance is stable on 8 consequent groups.

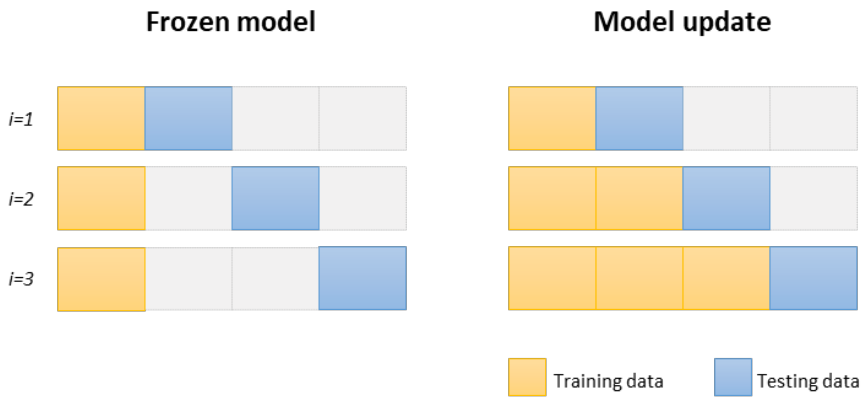

Figure 1. Schematic representation of the experiments with a frozen model and model update scenario. The frozen model was kept unchanged for all iterations while the model update was re-trained in every new iteration.

There are various types of control charts. Zone charts (or pre-control charts) are attractive due to their simple interpretation and were used to analyse the model's stability. Zone charts divide the chart into three zone limits in a “traffic light” design. The green zone, defined by mean ± 2 SDs of the process, indicates a stable process. The yellow zone, within 2-4 SDs of the mean, indicates a stable process if no two or more consecutive points fall in this zone. The points are the evaluated performance metrics (AUC and BS) used to assess the model's stability. The red zone, >4 SDs, indicates an unstable process if any point falls in this zone. All statistical analysis were performed with Python (version 3.8.8). To implement the models, the scikit-learn (version 0.24.1) and XGBoost (version 1.3.3) libraries were used.

Results

In total, data from 12,440 TAVI patients matched our inclusion period 2013-2019 and were considered for this study. For the analysis, data from 11,291 patients were included after excluding 837 patients for not having 30-day mortality information, 309 for belonging to the two centres with a high missing rate of mortality information, and 3 patients for having an additional procedure (i.e. not isolated TAVI).

The mean age of the included patients was 79.72 ± 6.86 and 50.21% of the patients were female. The baseline and procedural characteristics of the population used in this study, as well as the descriptive statistics, can be found in Table 1.

Table 1. Characteristics of the 11,291 TAVI patients stratified by their 30-day mortality survival status. Values are represented as mean and standard deviation (SD), median and interquartile range (IQR), number (n), or percentage (%).

		Grouped by 30-day mortality			
		Missing	**Non-surv.**	**Surv.**	**p-value**
n			410	10881	
Age (yr), mean (SD)		0	80.3 (7.2)	79.7 (6.9)	0.095
Sex, n (%)	*Male*	0	191 (46.6)	5,430 (49.9)	0.205
	Female		219 (53.4)	5,451 (50.1)	
BMI (kg/m^2), mean (SD)		143	26.5 (5.6)	27.3 (4.9)	0.010
Year of procedure, n (%)	*2013*	0	60 (14.6)	723 (6.6)	<0.001
	2014		61 (14.9)	973 (8.9)	
	2015		55 (13.4)	1,305 (12.0)	
	2016		57 (13.9)	1,450 (13.3)	
	2017		60 (14.6)	1,898 (17.4)	
	2018		60 (14.6)	2,073 (19.1)	
	2019		57 (13.9)	2,459 (22.6)	
eGFR (mL/min/1.73m2), mean (SD)		36	55.3 (21.6)	60.5 (29.3)	<0.001
NYHA class, n (%)	*1*	1151	32 (8.8)	1,067 (10.9)	<0.001
	2		62 (17.1)	2,718 (27.8)	
	3		216 (59.7)	5,377 (55.0)	
	4		52 (14.4)	616 (6.3)	
Chronic lung disease, n (%)	*No*	37	298 (73.8)	8,627 (79.5)	0.006
	Yes		106 (26.2)	2,223 (20.5)	
Procedure acuity, n (%)	*Elective*	220	312 (80.6)	9,733 (91.1)	<0.001
	Emergency		4 (1.0)	19 (0.2)	
	Urgent		71 (18.3)	932 (8.7)	
Dialysis, n (%)	*No*	212	387 (97.7)	10,570 (98.9)	0.042
	Yes		9 (2.3)	113 (1.1)	
TAVI access route, n (%)	*Direct aortic access*	15	56 (13.7)	781 (7.2)	<0.001
	Other access		2 (0.5)	15 (0.1)	
	Subclavian access		31 (7.6)	623 (5.7)	
	Transapical		53 (13.0)	679 (6.2)	
	Transf., other		37 (9.1)	684 (6.3)	

	Transf., percutaneous		179 (43.9)	6,104 (56.2)	
	Trasnf., surgical		50 (12.3)	1,982 (18.2)	
Critical preoperative state, n (%)	*No*	94	382 (96.2)	10,747 (99.5)	<0.001
	Yes		15 (3.8)	53 (0.5)	
Systolic pulmonary arterial pressure (mmHg), median (IQR)		2084	25.0 [25.0,38.5]	25.0 [25.0,31.2]	0.001

Non-surv: Non-survival, Surv: Survival, BMI: Body Mass Index, NYHA: New York Heart Association Functional Classification, TAVI: Transcatheter Aortic Valve Implantation, SD: Standard Deviation, IQR: Interquartile Range.

Internal validation

In the internal validation, with the inclusion of all 11,291 patients at the same time and a 10-fold CV, both the LR and XGB achieved a mean AUC of 0.68 and, respectively, a mean BS of 0.034 and 0.036 (Table 2). The calibration plots are available in the supplementary material Figure 1.

Group analysis

In Figure 2, the 30-day mortality and age of the patients are plotted over time. They demonstrate downward trends. When preparing data for the temporal validation for SPC analysis, we observed that the first 4 points (2013-2014) showed a trend, and were hence excluded (supplementary material Figure 2). The subsequent 8 groups did show stable performance and hence were used to train the frozen model and the initial model that will subsequently be updated.

Table 2. Evaluation of the models trained without temporal assessment (internal validation) with standard deviation. AUC = Area Under the Receiver Operating Characteristic curve, BS = Brier Score.

Model/Metric	AUC	BS
Logistic Regression	0.68 ± 0.07	0.034 ± 0.001
XGBoost	0.68 ± 0.05	0.036 ± 0.001

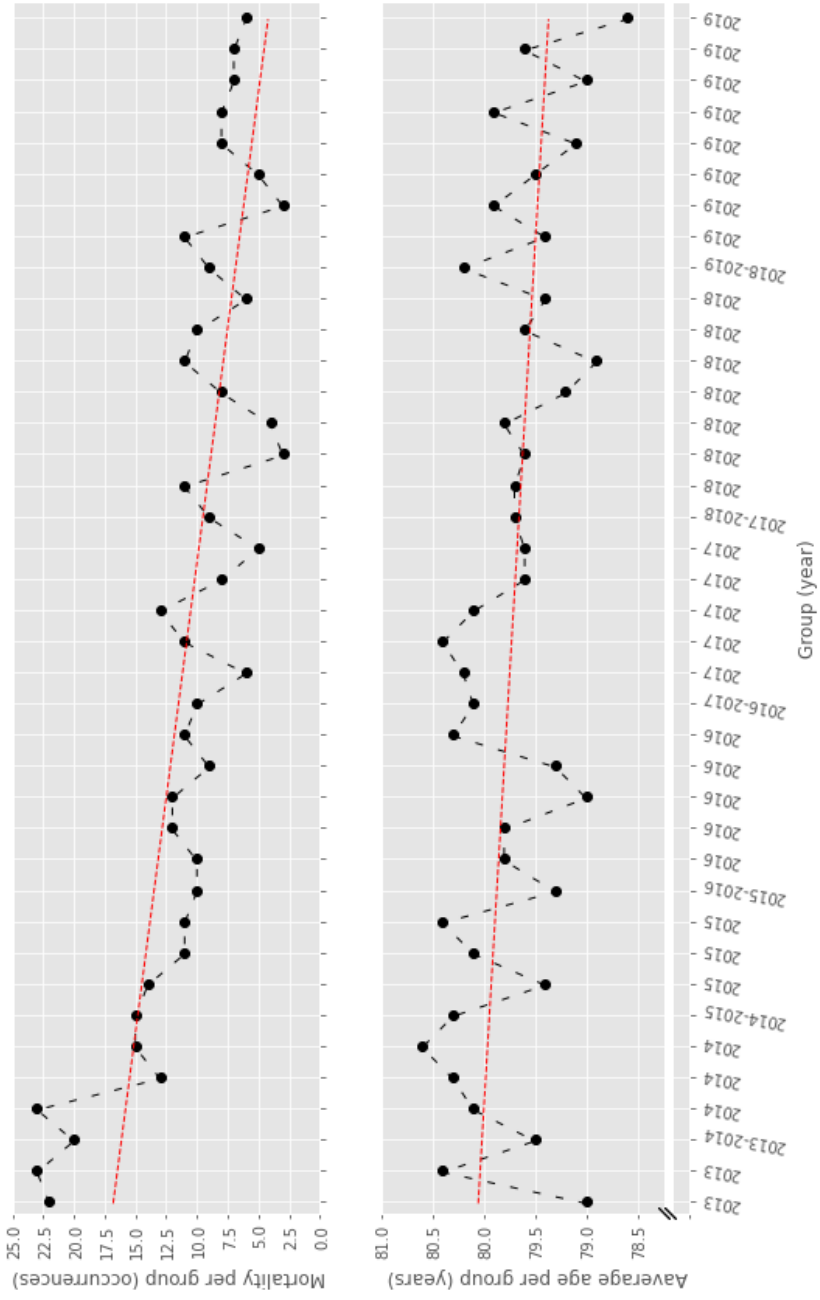

Figure 2. Mean 30-day mortality and age over time of TAVI patients. A linear trend is presented in red.

Statistical process control

Figure 3 displays the performance of the LR and XGB frozen models. While the AUC was considerably stable for the LR model and remained stable after 2017 for the XGB, BS was mostly in the red zone (> 4 SD) for both models. Figure 4 displays the progress of performance over time when using the model update approach (note that the zone limits are continuously updated as well). Both LR and XGB models were stable in their AUC, but instability in BS is observed at the beginning. The AUC limits of the updated LR model slightly changed compared to the frozen model. The AUC of the XGB model and BS of both models had their range visibly increased. This indicates a larger standard deviation which reflects larger uncertainty detected over time. The frozen parametric model had a median AUC of 0.64 (IQR 0.54-0.73) and BS of 0.028 (IQR 0.021-0.035) while the frozen non-parametric model had a median AUC of 0.63 (IQR 0.48-0.68) and BS of 0.027 (IQR 0.021-0.036). Regarding the model update, the parametric model had a median AUC of 0.66 (IQR 0.57-0.73) and BS of 0.027 (IQR 0.020-0.035) while the non-parametric had a median AUC of 0.66 (IQR 0.57-0.74) and BS of 0.027 (IQR 0.023-0.035).

Figure 5 shows the calibration curves (all points combined), with the frozen and model update approaches, for both LR and XGB models. The frozen LR and XGB models are completely overestimating the predicted mortality risk. The model update approach does achieve a more balanced calibration. The calibration plots, assessed over time, are available in the supplementary material Figure 3 (for LR) and supplementary material Figure 4 for (XGB).

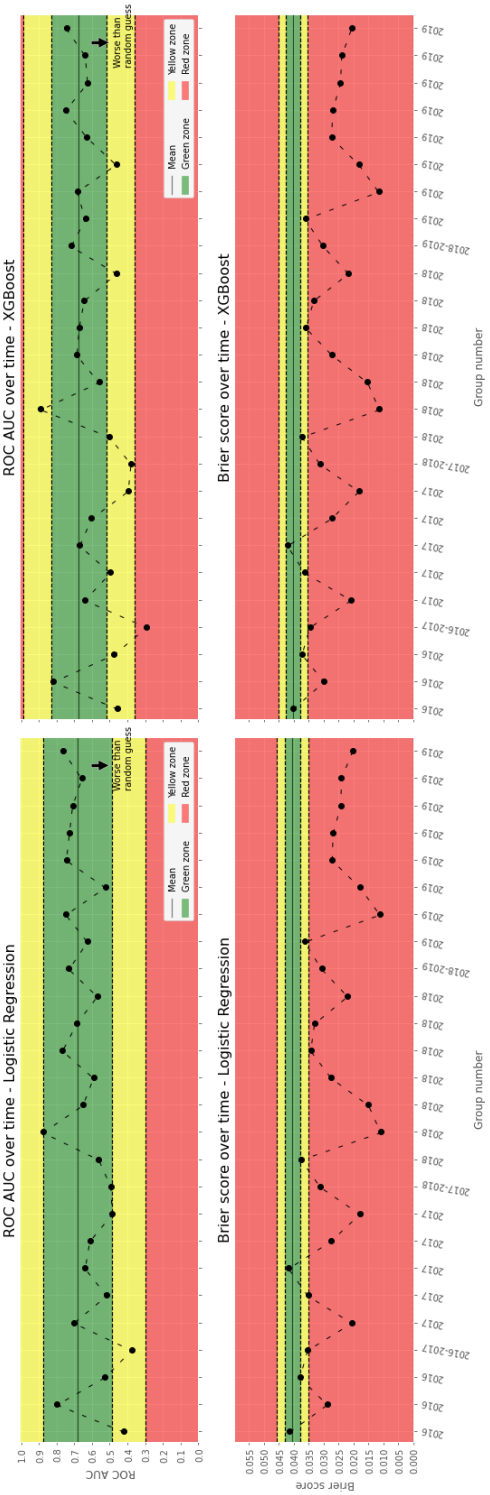

Figure 3. Temporal validation of the frozen LR (left) and XGB (right) models. For the LR model, the AUC is considered stable and most of the BS points are inside the red zone, hence the BS is unstable. Regarding the XGB model, an AUC point and most of the BS points are inside the red zone, hence the process is unstable.

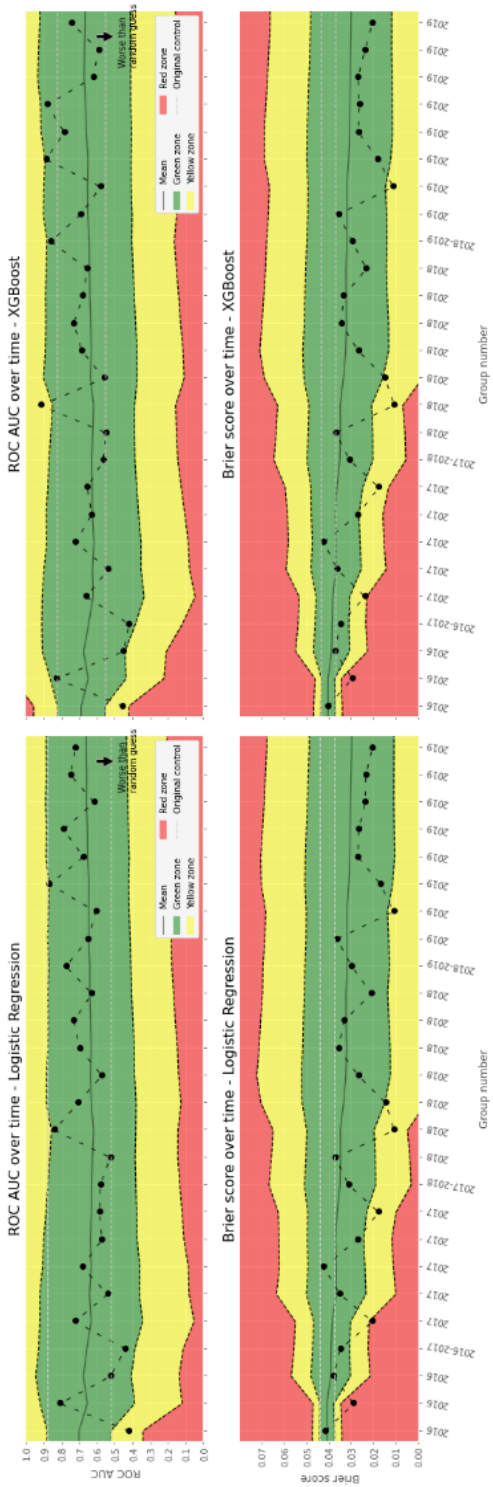

Figure 4. Temporal validation of the model update for LR (left) and XGB (right). While the AUC is stable for both models, some of the BS points are in the red zone at the beginning, hence the BS is unstable. Note that the zone limits are recalculated per update on a successive group.

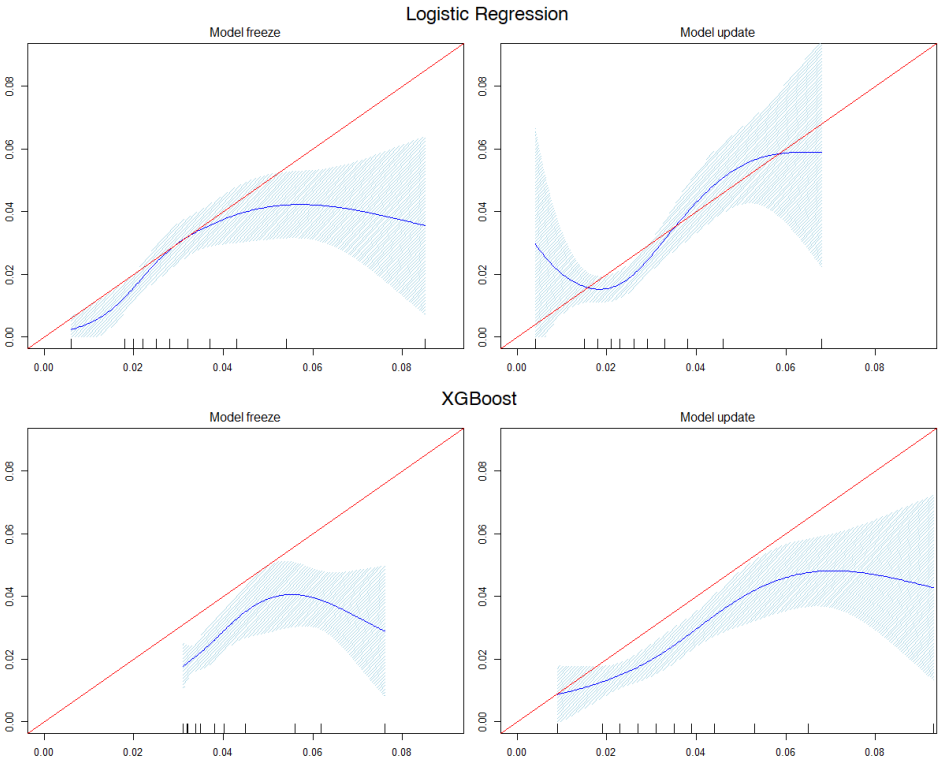

Figure 5. Calibration plots of the LR and XGB models. The plots were generated after the combination of all data points.

Discussion

Without repeated updates over time, the parametric TAVI mortality prediction model was considered stable regarding discrimination (in terms of the AUC) but unstable regarding the accuracy of the predicted probabilities (in terms of the BS). The non-parametric model was unstable in both AUC and BS. When models were repeatedly updated over time, both parametric and non-parametric models were considerably more stable and had only few points in the yellow and red zones.

TAVI procedures are over time offered to younger patients and patients with lower risk. Therefore, the mortality outcomes improved over time and the average age and mortality of the analysed TAVI population declined over time. When not updated, this decline in mortality results in the tendency of the prediction models to overestimate the mortality probability, which in turn leads to unstable models.

When we performed the model update analysis, which also updates the limits, a widening in these limits was clearly visible reflecting the larger uncertainty because of the higher variation in the data. It is important to note that the AUC is much less affected by the mortality prevalence than the BS. This explains why the BS become quickly unstable in the frozen models.

Regarding TAVI mortality prediction models, Al-Farra et al. (18) performed a prospective analysis of mortality prediction models and highlighted the importance of performing model updates to overcome performance drifts. In this latter study, two parametric models were analysed and the prospective data was treated as a single dataset, while we divided the prospective data into multiple groups and we used SPC. Also, recent studies using national registries from Germany and Switzerland (27,28) analysed temporal trends over the TAVI procedures, confirming the reduction in mortality we found. However, the accuracy and stability of the risk scores over time was not considered in these studies, nor was SPC used. Using SPC to investigate stability over time on prediction models had been used by Minne et al. (25,26) for evaluating pre-existent models for the prediction of mortality in the intensive care unit. Similar to our results, they found a significant difference within BS over time, while the AUC remained stable. However, they did not find time trends in the mortality or age of the observed patients. Also, the authors used a first-level recalibration approach, instead of re-training the model, to deal with the effects of time on the data. Although effective in their study, it was also suggested that more rigorous approaches, such as the model update we used, might be needed.

Strengths of this study include the use of a large national registry with more than 10,000 patients, with real recent data over many years. In addition, we compared two methods: a parametric (LR) and a non-parametric method (XGB). Furthermore, instead of simply analysing a frozen model with prospective data, we proposed a model update approach and evaluated its performance. Finally, we used two important performance measures: AUC for gauging discrimination and the BS for measuring the accuracy of the predicted probability. We also looked at the implications of (in)stability in terms of calibration graphs. As far as we know, this is the first study performing temporal analysis of TAVI mortality prediction models with such techniques.

This study also has some limitations. This data is from a national registry and has multiple centres. The centres might have different standards for patient selection or the performance of the procedure and this information was not taken into account directly (the centre was not used as a feature). In addition, the analysed data is from a country with a mainly Caucasian population, so a country-specific analysis. Also, a fixed number of samples per group was used to better understand how the models change over time and this leads to a different number of groups per year/month. Considering the clinical implementation of this study, one would have to wait until the number of procedures is reached to include a new group.

Our work shows the importance of taking time into account when using mortality prediction models. Specifically, in our large dataset, the stability of both parametric and non-parametric models was considered poor, mainly for the BS. This demonstrates the danger of only considering AUC when evaluating prediction models, which is a common practice, and the importance of analysing multiple metrics when evaluating models. With the model update, the stability increased for both parametric and non-parametric models. However, this improved stability came at the cost of more uncertainty in performance. We found that it might be risky to use a model for longer periods without updating, independent of whether it is a parametric or non-parametric model. The frozen models were poorly calibrated and, also with the model update, the calibration was still insufficient. An underestimation or overestimation of the predicted probability is seen in the calibration plots for both models. Finally, we would like to highlight the importance of inspecting the confidence intervals (reflecting honesty in uncertainty) rather than the absolute improvement of the models' performance.

Future work can investigate differences between centres. Also, in order to avoid the necessity of having enough patients to compose a new group. Individual Control Charts, that are able to analyse individual measurements, could be explored. In addition, a subgroup analysis could provide insight into specific groups that markedly diverge from the rest of the population. Furthermore, one hypothesis is that the older data might harm the model and can be ignored, or given less weight, over time once it might be too different from the current

population. Finally, reproducing this experiment with a different (TAVI) population warrants further research.

Conclusion

In our study, the prediction models that were updated over time were more stable and accurate compared to the frozen models. It highlights the importance of repeatedly updating the models over time to improve their performance stability. Although the updated models were more stable, the calibration was still poor and it came also at the cost of more uncertainty in performance. There were no clear benefits in using the non-parametric model over the parametric model. The trained models, when not updated, were unstable and presented a higher overestimation of 30-day mortality after TAVI than the models that were updated over time.

References

1. Carabello BA, Paulus WJ. Aortic stenosis. Lancet. 2009;373(9667):956–66.
2. Nielsen HH. Transcatheter aortic valve implantation. Dan Med J. 2012;59(12):B4556.
3. Fang F, Tang J, Zhao Y, He J, Xu P, Faramand A. Transcatheter aortic valve implantation versus surgical aortic valve replacement in patients at low and intermediate risk: A risk specific meta-analysis of randomized controlled trials. PLoS One. 2019;14(9):e0221922.
4. Gomez CA, Braghiroli J, de Marchena E. "The changing paradigm": TAVR for low-risk patients approved by the FDA. Wiley Online Library; 2020.
5. Akodad M, Lefèvre T. TAVI: Simplification Is the Ultimate Sophistication. Vol. 5, Frontiers in Cardiovascular Medicine. 2018. p. 96.
6. Abdelaziz HK, Roberts DH. Advances in transcatheter aortic valve implantation. In: Ahmed W, Phoenix DA, Jackson MJ, Charalambous CPBT-A in M and SE, editors. Academic Press; 2020. p. 103–19.
7. Nishimura RA, Otto CM, Bonow RO, Carabello BA, Erwin JP, Guyton RA, et al. 2014 AHA/ACC guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63(22):e57–185.
8. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 2-Isolated Valve Surgery. Ann Thorac Surg [Internet]. 2009;88(1 SUPPL.):S23–42. Available from: http://dx.doi.org/10.1016/j.athoracsur.2009.05.056
9. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R, et al. European system for cardiac operative risk evaluation (Euro SCORE). Eur J cardio-thoracic Surg. 1999;16(1):9–13.
10. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. Euroscore ii. Eur J cardio-thoracic Surg. 2012;41(4):734–45.
11. Iung B, Laouénan C, Himbert D, Eltchaninoff H, Chevreul K, Donzeau-Gouge P, et al. Predictive factors of early mortality after transcatheter aortic valve implantation: individual risk assessment using a simple score. Heart. 2014;100(13):1016–23.
12. Edwards FH, Cohen DJ, O'Brien SM, Peterson ED, Mack MJ, Shahian DM, et al. Development and validation of a risk prediction model for in-hospital mortality after transcatheter aortic valve replacement. JAMA Cardiol. 2016;1(1):46–52.
13. Lopes RR, van Mourik MS, Schaft E V., Ramos LA, Baan J, Vendrik J, et al. Value of machine learning in predicting TAVI outcomes. Netherlands Hear J [Internet]. 2019 Sep 20;27(9):443–50. Available from: http://link.springer.com/10.1007/s12471-019-1285-7
14. Agasthi P, Ashraf H, Pujari SH, Girardo ME, Tseng A, Mookadam F, et al. Artificial intelligence trumps TAVI2-SCORE and CoreValve Score in predicting 1-year mortality post Transcatheter Aortic Valve Replacement. Cardiovasc Revascularization Med. 2020;
15. Al-Farra H, Abu-Hanna A, de Mol BAJM, Ter Burg WJ, Houterman S, Henriques JPS, et al. External validation of existing prediction models of 30-day mortality

after Transcatheter Aortic Valve Implantation (TAVI) in the Netherlands Heart Registration. Int J Cardiol. 2020;317:25–32.
16. Wolff G, Shamekhi J, Al-Kassou B, Tabata N, Parco C, Klein K, et al. Risk modeling in transcatheter aortic valve replacement remains unsolved: an external validation study in 2946 German patients. Clin Res Cardiol. 2020;1–9.
17. Martin GP, Sperrin M, Ludman PF, de Belder MA, Gale CP, Toff WD, et al. Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation. Am Heart J. 2017;184:97–105.
18. Al-Farra H, de Mol BAJM, Ravelli ACJ, Ter Burg W, Houterman S, Henriques JPS, et al. Update and, internal and temporal-validation of the FRANCE-2 and ACC-TAVI early-mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) using data from the Netherlands heart registration (NHR). IJC Hear Vasc. 2021;32:100716.
19. Lopes RR, Mamprin M, Zelis JM, Tonino PAL, van Mourik MS, Vis MM, et al. Inter-Center Cross-Validation and Finetuning without Patient Data Sharing for Predicting Transcatheter Aortic Valve Implantation Outcome. 2020 IEEE 33rd Int Symp Comput Med Syst [Internet]. 2020 Jul;591–6. Available from: https://ieeexplore.ieee.org/document/9183069/
20. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Statistical process control for validating a classification tree model for predicting mortality–a novel approach towards temporal validation. J Biomed Inform. 2012;45(1):37–44.
21. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Statistical process control for monitoring standardized mortality ratios of a classification tree model. Methods Inf Med. 2012;51(4):353–8.
22. Timmermans MJC, Houterman S, Daeter ED, Danse PW, Li WW, Lipsic · Erik, et al. Using real-world data to monitor and improve quality of care in coronary artery disease: results from the Netherlands Heart Registration. Netherlands Hear J 2022 [Internet]. 2022 Apr 7 [cited 2022 Apr 24];1–9. Available from: https://link.springer.com/article/10.1007/s12471-022-01672-0
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012;12:2825–30.
24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016;785–94.
25. Minne L, Eslami S, De Keizer N, De Jonge E, De Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med [Internet]. 2012 Jan [cited 2022 Mar 28];38(1):40–6. Available from: https://pubmed.ncbi.nlm.nih.gov/22042520/
26. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Statistical process control for monitoring standardized mortality ratios of a classification tree model. Methods Inf Med [Internet]. 2012 [cited 2022 Mar 28];51(4):353–8. Available from: https://pubmed.ncbi.nlm.nih.gov/22773038/
27. Mauri V, Abdel-Wahab M, Bleiziffer S, Veulemans V, Sedaghat A, Adam M, et al. Temporal trends of TAVI treatment characteristics in high volume centers in Germany 2013–2020. Clin Res Cardiol [Internet]. 2021 Nov 9 [cited 2022 Mar 10];1–8. Available from: https://link.springer.com/article/10.1007/s00392-021-01963-3

28. Stortecky S, Franzone A, Heg D, Tueller D, Noble S, Pilgrim T, et al. Temporal trends in adoption and outcomes of transcatheter aortic valve implantation: a SwissTAVI Registry analysis. Eur Hear J - Qual Care Clin Outcomes [Internet]. 2019 Jul 1 [cited 2022 Mar 10];5(3):242–51. Available from: https://academic.oup.com/ehjqcco/article/5/3/242/5124351

Supplementary Material

Table S1. Hyperparameters grid used for XGB and LR.

Classifier	Parameter name	Parameter value
XGB	Number of trees	[1000]
	Early stopping rounds	[10]
	Validation set	[0.1]
	Subsample	[0.9, 0.7]
	Max depth	[2, 4, 8]
	Min child weight	[1, 5]
	Gamma	[0, 1, 5, 10]
	Colsample by tree	[1, 0.7]
	Learning rate	[0.1, 0.05]
	Scale pos weight	[1, 2, 3]
LR	Penalty	[L1, L2]
	Solver	[liblinear, lbfgs]
	C	[0.1, 1, 10]

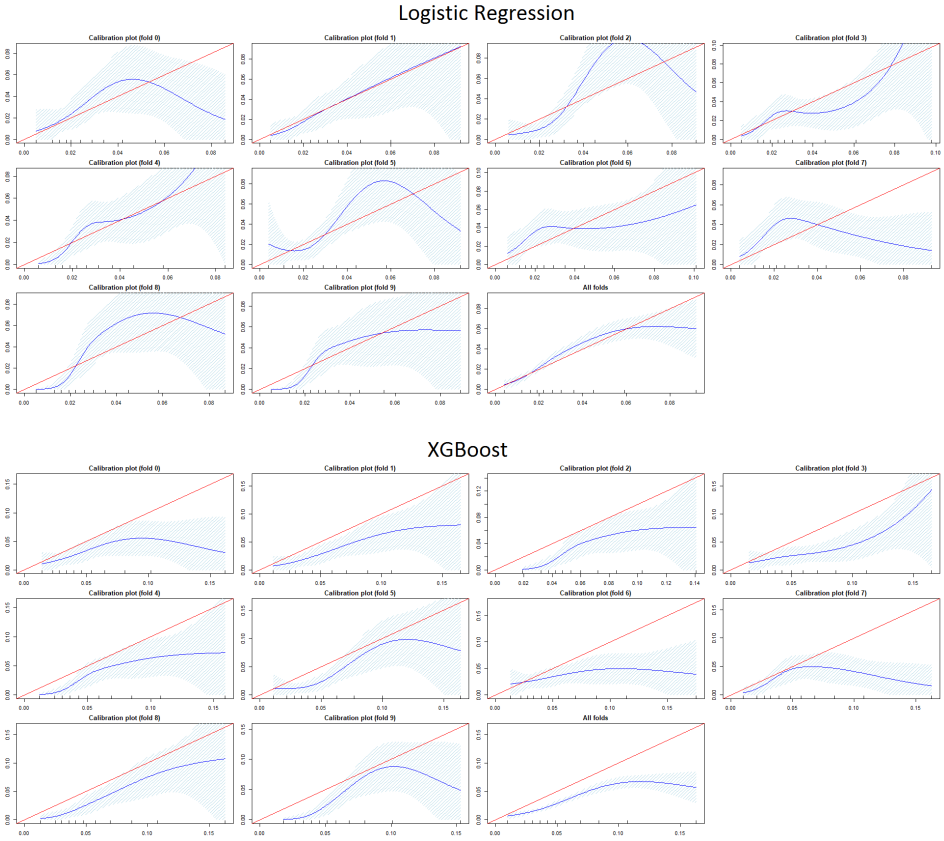

Figure 1. Calibration plots for LR and XGB (internal validation) per fold and all combined.

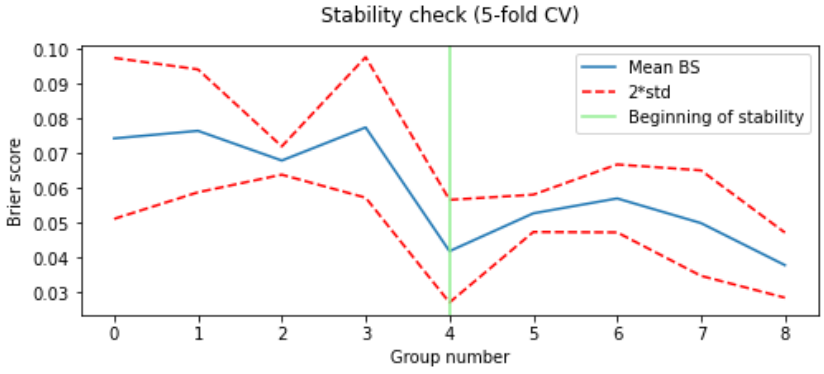

Figure 2. Stability check using Logistic Regression for the initial points. The first 4 points were excluded since they were not considered stable

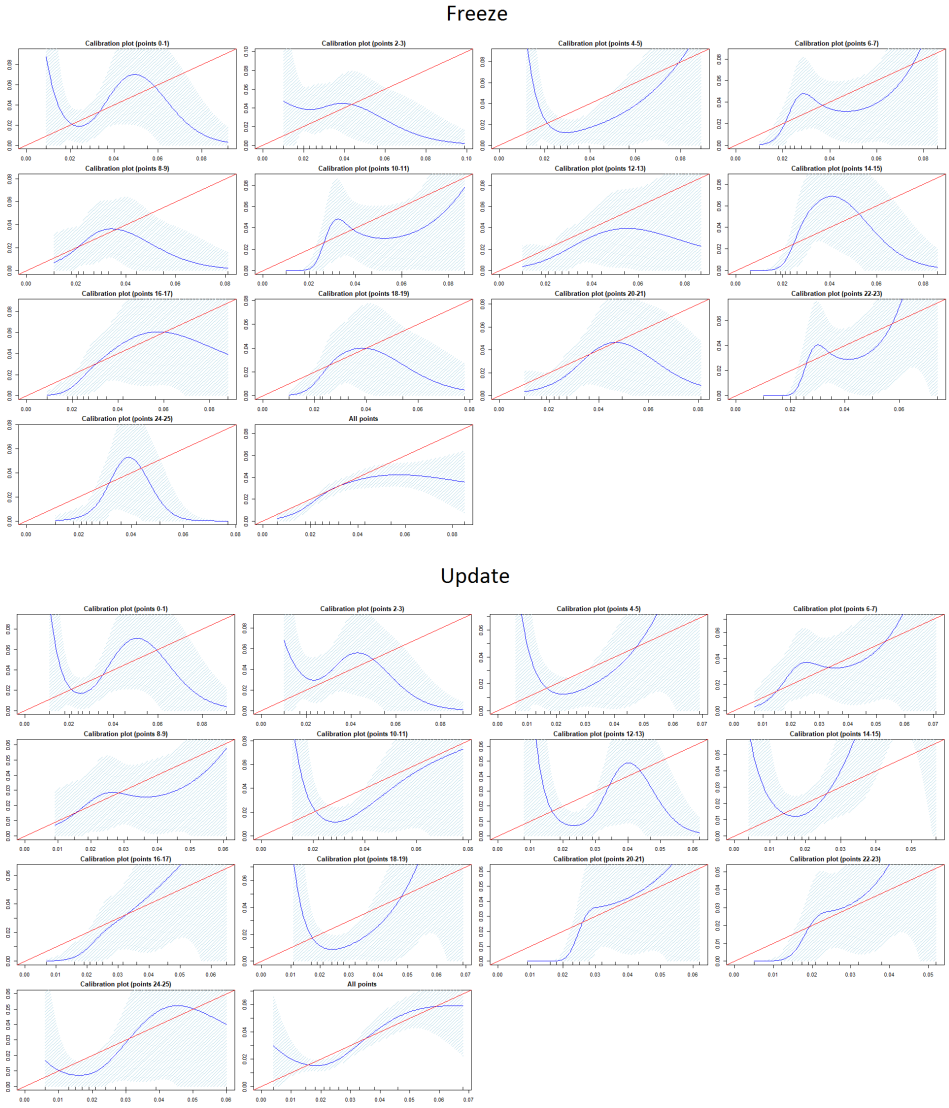

Figure 3. Calibration plots for LR (temporal validation) per 2 groups and all combined.

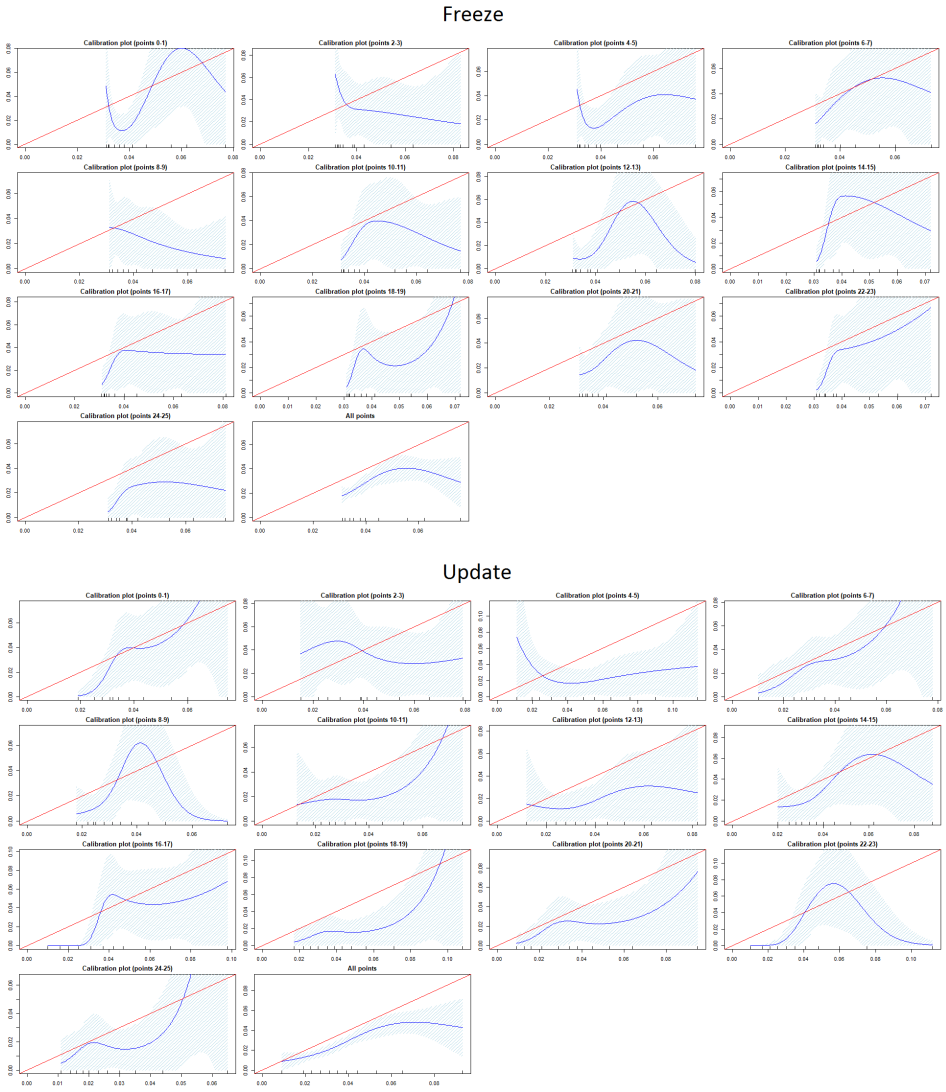

Figure 4. Calibration plots for XGB (temporal validation) per 2 groups and all combined.

6

Prediction of atrial fibrillation recurrence after thoracoscopic surgical ablation using machine learning techniques

Baalman SW*, Lopes RR*, Ramos LA, Neefs J, Driessen AH, van Boven WP, de Mol BA, Marquering HA, de Groot JR.

Diagnostics. 2021 Oct;11(10):1787.

* Shared first author

DOI: 10.3390/diagnostics11101787

Abstract

Thoracoscopic surgical ablation (SA) for atrial fibrillation (AF) has shown to be an effective treatment to restore sinus rhythm in patients with advanced AF. Identifying patients who will not benefit from this procedure would be valuable to improve personalized AF therapy. Machine learning (ML) techniques may assist in the improvement of clinical prediction models for patient selection. The aim of this study is to investigate how available baseline characteristics predict AF recurrence after SA using ML techniques.

One-hundred-sixty clinical baseline variables were collected from 446 AF patients undergoing SA in our tertiary referral center. Multiple ML models were trained on five outcome measurements, including either all or a number of key variables selected by using the least absolute shrinkage and selection operator (LASSO).

There was no difference in model performance between different ML techniques or outcome measurements. Variable selection significantly improved model performance (AUC: 0.73, 95% CI: 0.68–0.77). Subgroup analysis showed a higher model performance in younger patients (< 55 years, AUC: 0.82 vs. > 55 years, AUC 0.66). Recurrences of AF after SA can be predicted best when using a selection of baseline characteristics, particularly in young patients.

Introduction

In patients with advanced atrial fibrillation (AF), thoracoscopic surgical ablation (SA) is effective to restore sinus rhythm (SR) (1). Minimally invasive SA for AF using video-assisted thoracoscopic surgery has increasingly been performed and has a success rate of 69–80% in terms of freedom of AF at one year after surgery (2).

Several clinical variables predicting AF recurrence after catheter ablation (CA) have been identified. These variables are currently being applied for patient selection for both CA and SA (3). Despite our knowledge of risk factors that are associated with lower efficacy and more recurrences, there are no risk scores or prediction models available that consider all the available pre-procedural clinical data that may affect the outcome of SA. More importantly, it is unknown to what extent the AF recurrence risk after an SA procedure is embedded in baseline clinical characteristics, and to what extent the AF recurrence risk is purely stochastic or related to technical aspects of the procedure (i.e., reconnection across ablation lines). Therefore, a systematic analysis tool to assess the risk of any ablation failure could potentially lead to enhanced identification of patients who may benefit from SA versus those in whom SA therapy would be futile.

Conventionally developed clinical prediction models are using traditional linear regression methods. As an alternative, other machine learning (ML) techniques enable the discovery of (novel) potentially complex patterns in data sets through automated algorithms, using techniques like the kernel trick or multilayer neural network, which may result in more efficient processing of non-linear relationships and complex interactions between variables (4). ML has already been successfully used on many studies enabling the detection and diagnosis of AF (5). By using a ML approach, more effective selection and weighing of parameters of choice can be achieved, leading to promising clinical prediction models, which may be more accurate than classical prediction models (6,7). Still, the ultimate predictive value of such models will depend on the proportion of risk factors present in the variables that are causally related to an outcome versus non-predictive risk factors that are randomly distributed among subjects. Therefore, we sought to optimize the prediction of AF recurrence following SA with the use

of available clinical, laboratory and imaging data to investigate to what extent the risk of AF recurrence is already embedded in the preoperative data.

In this study, we built several ML models that incorporate preoperative data in AF patients scheduled for SA to comprehensively predict the AF recurrence risk. The aim of this study was I) to evaluate the proportion of baseline characteristics that are causal risk factors for AF recurrence after SA using different ML techniques; II) to investigate the differential performance of ML models on multiple conventional and modified definitions of AF recurrence; and III) to analyze whether the accuracy of the ML models is pertinent for clinically relevant subgroups.

Materials and Methods

Patient Characteristics

Patients with paroxysmal or persistent AF who underwent SA in our center between February 2008 to June 2017 were eligible for this analysis. All patients provided written informed consent before the procedure. Clinical variables collected prior to SA were used for further analysis and consisted of patients' characteristics, AF type and duration, medical history, the (determinants of the) CHA_2DS_2-VASc score, medication, Holter and electrocardiogram (ECG) reports, vital parameters, imaging (i.e., echocardiography, magnetic resonance imaging, computer tomography), and laboratory measurements. A full list of all collected variables is shown in Supplementary Material Table I. All continuous variables were standardized by removing the mean and scaling to unit variance. For categorical variables we used one-hot encoding (also known as "dummy coding").

Procedure and Outcome

Included patients underwent SA following our standard protocol, using a hybrid surgical–electrophysiological approach as described previously (8,9). Approximately half of the patients underwent additional ganglion plexus (GP) ablation as part of the standard of care in all procedures performed before 2010, or as part of participation in the randomized Atrial Fibrillation Ablation and Autonomic Modulation via Thoracoscopic Surgery (AFACT) trial (2). As the

AFACT trial demonstrated, there was no difference in AF recurrence between the randomized treatment groups, so data of patients with and without GP ablation were pooled. Patients were followed for 24 months after SA with frequent ECG and 24 h-Holter monitoring (2).

Five different definitions of AF recurrence were applied:

- Outcome 1: any episode of atrial tachyarrhythmia (AF, atrial flutter, atrial tachycardia) lasting > 30 s (10).
- Outcome 2: any episode of AF (but not atrial flutter or atrial tachycardia) lasting > 30 s.
- Outcome 3: one single episode of any atrial tachyarrhythmia lasting > 1 h.
- Outcome 4: one single episode of any atrial tachyarrhythmia lasting > 6 h.
- Outcome 5: one single episode of AF (but not atrial flutter or atrial tachycardia), lasting > 1 h.

All outcomes were assessed during the two-year follow-up period, with exclusion of the first three months following the procedure, which were considered a blanking period for outcome analysis.

Missing Data

Missing data was imputed with MissForest (11), which is an iterative imputation method based on random forest. Only the training set was used to train the imputation model. The target variables (different definition of AF recurrences) were not included in this process. Variables that were less than 70% complete and patients with more than 70% missing data were, sequentially, discarded from the analysis.

Machine Learning Algorithms

Five well-established ML algorithms were selected: support vector machine (SVM), logistic regression (LR), random forest (RF), neural network (NN), and gradient boosting (GB). All models were implemented using scikit-learn (12). Furthermore, we applied the least absolute shrinkage and selection operator (LASSO), which performs a regularization to automatically select variables and reduces the number of variables by fitting a linear regression with L1 regularization. This is done to decrease the model's complexity and reduce the

input noise (13). Variable selection steps are expected to reduce redundant or irrelevant data and can lead to an increase in the model's accuracy (14).

Analysis Pipeline and Variable Selection

A nested cross-validation (CV), with an internal and external CV, was used for evaluation. The external CV was a stratified 5-fold, which means that 80% of the data was used for training and 20% for testing (repeated five times until all data is used for both training and test). The test set was not used during training and validation steps.

The internal CV, also a stratified 5-fold, was first used by LASSO to select the variables to assure that the model generalized well to different data samples. Variables selected more than once in the CV by LASSO were subsequently included to train the models (13). This strategy was adopted to avoid the chance of selecting a variable that was only meaningful to predict a single fold. Subsequently, the same internal 5-fold CV was used to determine the best hyperparameters by grid search for each classifier on each fold and to train the models. The hyperparameter ranges used are displayed in Supplementary Material Table II and III. The pipeline, shown in Figure 1, was ran for all the outcome measurements as target variables.

Model Evaluation

The area under the curve (AUC) of the receiver operating characteristic was used to evaluate the performance of each model (external CV) and to select the model after the hyperparameter optimization (internal CV). Since a 5-fold CV was used for evaluation, we computed the mean AUC, standard deviation (SD) and confidence interval (CI) of each classifier.

Subgroup Analysis

We performed a predefined subgroup analysis using the model structure (outcome measurement, (key)-variables, ML algorithm) of the two best performing models. For this analysis, the probability prediction from the test sets (from all 5-folds) were combined, creating a single distribution with a single prediction probability for each sample. Samples were selected from this distribution given their subgroups and an AUC was computed for each subgroup

individually. Subgroups were chosen based on their established predictive value for AF recurrence or inclusion in the CHA_2DS_2-VASc score (15,16). Variables with an unbalanced distribution were not taken into account. The following variables were included for subgroup analysis: CHA_2DS_2-VASc score, congestive heart failure, history of stroke, history of CA, vascular disease, diabetes, hypertension, left atrial volume index (LAVI), sex, and age. Subgroups were created by using the predefined categories in case of categorical variables, and quartiles in case of continuous variables.

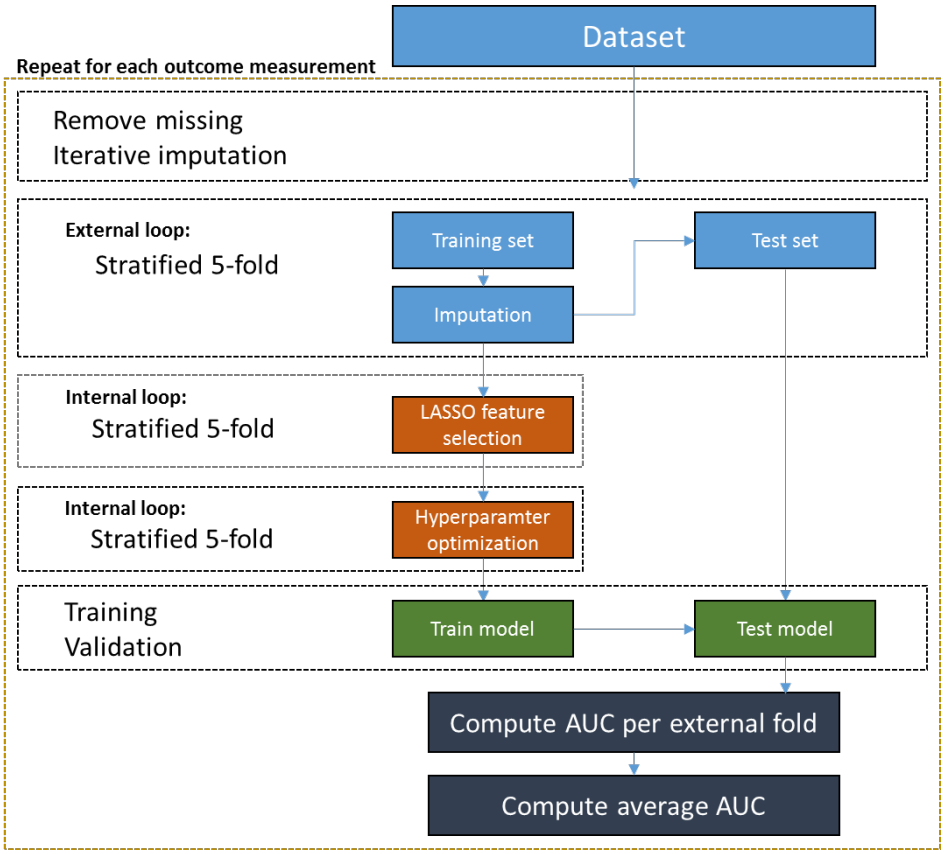

Figure 1. Schematic representation of nested cross-validation methodology. Initially, the missing data is removed and an iterative imputation is performed in a stratified 5-fold CV (external) using only the training set. The imputation model is further used to imput the test set. After that, an internal CV is performed for the LASSO feature selection and hyperparameter optimization. As the last step, the model is trained with the training set and validated with the test set. An average AUC is reported.

Model Interpretation

To increase the interpretability of our results, we explored the predictive impact of the selected features in our two best performing models. To gain more insight, we applied the unified framework Shapley additive explanations (SHAP) for the interpretation of predictions, which can be used for both linear and non-linear models (17). The SHAP was calculated for each feature comparing the prediction of the model without that feature. In addition, in cases where LR proved to be the best performing model, we used the coefficients of each feature to provide an interpretation of how each individual feature affected the prediction.

Statistical Analysis

Continuous data are presented as mean (SD) or median (range) for normally and non-normally distributed data, respectively. The unpaired T-test and Mann–Whitney U test were used for comparisons of AUCs between two groups. One-way ANOVA and Kruskal–Wallis tests were used for comparisons of AUCs between more than two groups. Statistical analyses were performed using SPSS Version 26 (IBM Corporation, Armonk, NY, United States). ML were developed with Python programming language 3.6 (Python Software Foundation, http://www.python.org/).

Results

Of the 495 patients, 49 (10%) patients were excluded because of incomplete baseline data. The mean age of the 446 included patients was 60 (SD ± 9) years, 335 (75%) were male and 266 (60%) had persistent AF (Table 1). An overview of baseline characteristics stratified by success or failure according to different outcome definitions is shown in Supplementary Material Table IV. In total, 18 out of 160 baseline variables (11%) were excluded because of missing values in more than 30% of the patients.

Prediction of AF Recurrence within Two Years after SA.

Outcome 1

A total of 188 (42%) of the 446 patients experienced recurrence of AF within two years after SA according to the definition of AF recurrence following current guidelines (Outcome 1). Prediction of AF recurrence, and all baseline

characteristics, resulted in an AUC varying from 0.53 (95% CI: 0.38–0.68 [SVM]) to 0.66 (95% CI: 0.59–0.72 [RF]) (Figure 2, Table 2). Variable selection using LASSO resulted in a selection of 12 key variables on the 5-fold CV. Variables regarding left atrial (LA) size, age and comorbidity (i.e., use of ACE inhibitors) demonstrated to be the most frequently (100%) selected variables to predict AF recurrence defined as Table 3. Training the models on Outcome 1 with the 12 selected key-variables resulted in an improved AUC up to 0.70 (95% CI: 0.62–0.78 (LR)).

Outcomes 2–5

In line with the results of Outcome 1, model performance significantly improved for all other outcome definitions using selected key variables instead of using all 142 available variables ($p < 0.001$). There were no significant differences in model performance between all outcome definitions ($p = 0.35$), nor in model performance between different ML techniques ($p = 0.28$). However, the best performing model for Outcome 2 (LASSO, LR) had a higher AUC (0.73, 95% CI: 0.68–0.77) compared to the best performing model of Outcome 1 (LASSO, LR; AUC: 0.70, 95% CI: 0.62–0.78). Figure 2 shows the average 5-fold ROC of model training for Outcome 1 and Outcome 2 with all and a selection of variables.

Table 1. Summarized patients' characteristics for all included patients.

Variable		No. of Patients (%)
n		*446*
Sex, *n* (%)		
	Male	335 (75.1)
	Female	111 (24.9)
BMI, mean (SD)		25.8 (7.5)
Age, mean (SD)		60.0 (8.7)
CHA_2DS_2-VASc, *n* (%)		
	0	122 (27.4)
	1	141 (31.6)
	≥ 2	183 (41.0)
AF type, *n* (%)		
	Paroxysmal	180 (40.4)
	Persistent	266 (59.6)

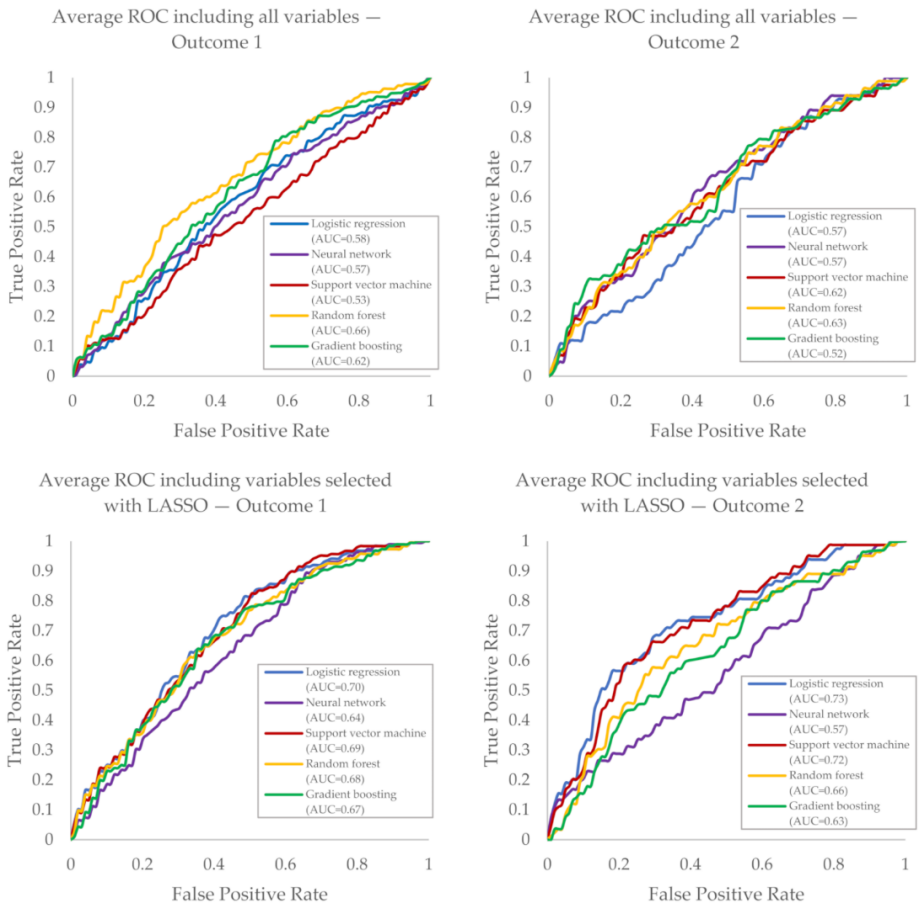

Figure 2. Average 5-fold ROC of testing on Outcome 1 or Outcome 2 with all or a selection of variables for two years without AF recurrence. ROC receiver operating characteristic curves, AUC area under the curve.

Table 2. Average area under the curve (AUC) and 95% confidence interval (CI).

Models, AUCs (95% CI)	Gradient Boosting	Random Forest	Support Vector Machine	Neural Network	Logistic Regression
Outcome 1	0.62 (0.55–0.68)	0.66 (0.59–0.72)	0.53 (0.38–0.68)	0.57 (0.44–0.71)	0.58 (0.54–0.61)
Outcome 1 with LASSO	0.67 (0.65–0.69)	0.68 (0.63–0.73)	0.69 (0.66–0.73)	0.64 (0.61–0.67)	0.70 (0.62–0.78)
Outcome 2	0.52 (0.48–0.57)	0.63 (0.57–0.69)	0.62 (0.55–0.70)	0.57 (0.50–0.64)	0.57 (0.50–0.64)
Outcome 2 with LASSO	0.63 (0.54–0.72)	0.66 (0.55–0.76)	0.72 (0.66–0.78)	0.57 (0.50–0.64)	0.73 (0.68–0.77)
Outcome 3	0.68 (0.59–0.76)	0.67 (0.61–0.72)	0.56 (0.47–0.65)	0.54 (0.42–0.65)	0.54 (0.48–0.60)
Outcome 3 with LASSO	0.67 (0.56–0.78)	0.69 (0.64–0.75)	0.67 (0.63–0.71)	0.68 (0.62–0.74)	0.69 (0.65–0.74)
Outcome 4	0.64 (0.52–0.75)	0.63 (0.55–0.72)	0.56 (0.43–0.69)	0.61 (0.57–0.64)	0.56 (0.48–0.63)
Outcome 4 with LASSO	0.65 (0.56–0.73)	0.67 (0.58–0.76)	0.66 (0.58–0.74)	0.62 (0.52–0.73)	0.68 (0.59–0.77)
Outcome 5	0.51 (0.43–0.59)	0.55 (0.51–0.59)	0.55 (0.42–0.67)	0.54 (0.37–0.70)	0.56 (0.51–0.62)
Outcome 5 with LASSO	0.63 (0.58–0.68)	0.67 (0.61–0.73)	0.66 (0.57–0.75)	0.55 (0.35–0.75)	0.69 (0.60–0.78)

Table 3. Key-variables for Outcome 1 and Outcome 2, ranked by the percentage the variable was selected during the 5-fold cross validation (1-fold = 20%). Variables selected by LASSO, in at least two folds (40%), were included for training the models.

Outcome 1		Outcome 2	
Variable - Assessment at Baseline	**Selection**	**Variable**	**Selection**
LAVI - *TTE*	100%	Max. SBP—*X-ECG*	100%
PR-interval - *ECG*	100%	ACE-inhibitor (use)—*medication*	100%
LA craniocaudal axis index - *CT*	100%	ARB (use) - *medication*	80%
Max. SBP - *X-ECG*	100%	LAVI - *TTE*	60%
ACE-inhibitor (use) - *medication*	100%	Total duration - *X-ECG*	60%
Age - *demographics*	100%	FVC - *lung capacity test*	60%
LA anteroposterior axis index - *CT*	80%	Class II antiarrhythmics (use) - *medication*	60%
Max. resistance - *X-ECG*	80%	Loop diuretics (dose) - *medication*	60%
Previous catheter ablation - *medical history*	80%	HR - *ECG*	60%
RSPV (width) - *CT*	60%	LA craniocaudal axis index - *CT*	40%
FEV1 - *lung capacity test*	60%	Previous catheter ablation - *medical history*	40%
Height - *physical examination*	60%	Total duration of AF - *Holter monitoring*	40%
Type of AF - *medical history*	60%		
Tricuspid valve regurgitation - *TTE*	40%		
FVC - *lung capacity test*	40%		
Hs-troponine - *blood sampling*	40%		
Class II antiarrhythmics (use)—*medication*	40%		
Class III antiarrhythmics (dose)—*medication*	40%		

AF atrial fibrillation, *ARB* angiotensin receptor blockers, *CT* computed tomography, *ECG* electrocardiogram, *FEV1* forced expiratory volume in one second, *FVC* forced vital capacity, *HR* heart rate, *HS-troponine* high sensitive troponine, *LA* left atrium, *LAVI* left atrial volume index, *RSPV* right superior pulmonary vein, *SBP* systolic blood pressure, *TTE* transthoracic echocardiography, *X-ECG* exercise testing

Model Interpretation Analysis

Feature importance (SHAP) of each key variable for the two best prediction models (LR, SVM) regarding Outcome 1 and Outcome 2 was calculated and averaged over the test folds (Figure 3). For both outcomes, the key variables with the highest SHAP values (amplitude) were consistent for the two models. For Outcome 1, AF type, maximal systolic blood pressure (SBP) during exercise testing, increased craniocaudal index of the LA on CT, and PR interval on the

baseline ECG were the key variables with the highest SHAP values. Hence, patients with persistent AF had a higher risk of AF recurrence (defined as Outcome 1) than patients with paroxysmal AF. In addition, for the continuous variables, the progressive change in color in Figure 3 indicates a possible linear relationship between the value of the variable and Outcome 1. Patients with a low maximal SBP during exercise testing, increased craniocaudal index of the LA and prolonged PR interval had a higher risk of AF recurrence (Outcome 1). For Outcome 2, maximal SBP during exercise testing, loop diuretics dose and heart rate on the baseline ECG were key variables with the highest SHAP values for both models. There was no difference in the direction of the SHAP values between the models of Outcome 1 and Outcome 2. As LR proved to be the best performing ML technique for both Outcome 1 and Outcome 2, we calculated the average LR coefficients (Supplementary Material Table V).

Analysis of AF Recurrence Prediction in Subgroups

Figure 4 shows the results of the balanced subgroups ranked by AUC for Outcome 1 and Outcome 2. There was an interaction between model performance and age, with the best performance of the model in patients < 55 years old (AUC: 0.82) for Outcome 2.

Discussion

This study of 446 patients undergoing SA for paroxysmal or persistent AF in our center aimed to improve patient selection for SA by investigating the value of baseline characteristics for the prediction of AF recurrence. Our main findings are: I) investigated ML models perform moderately well in the prediction of AF recurrence when all available baseline variables are included, but, with a selection of key variables, the prediction of AF recurrence improves; II) there are no differences in model performance using modified definitions of AF recurrence or different ML techniques; and III) subgroup analysis shows an improved model performance in younger patients.

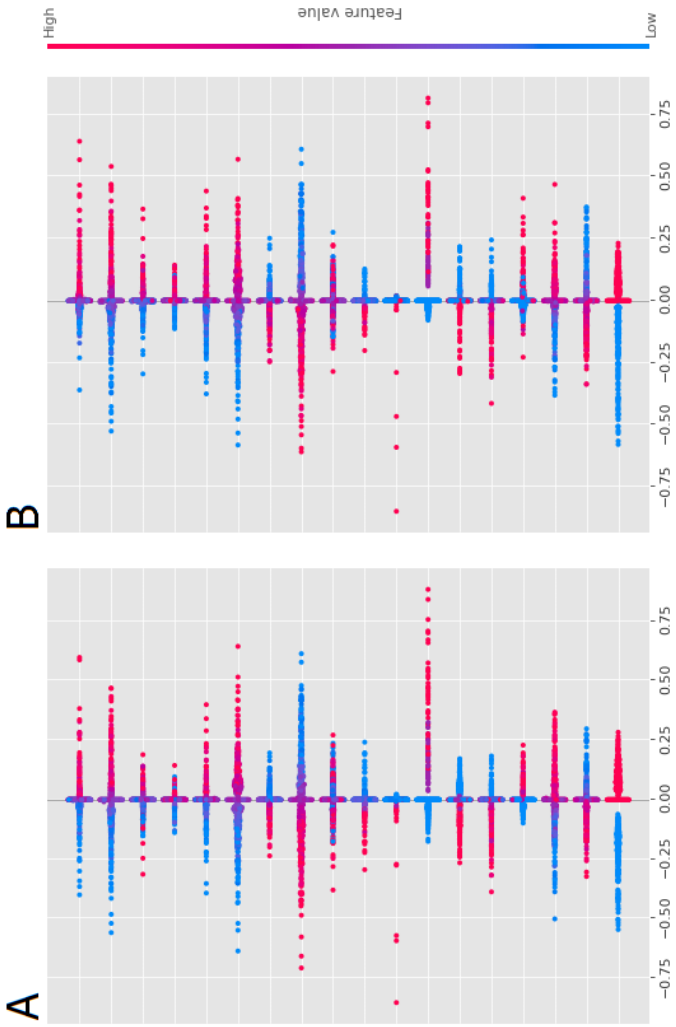

Figure 3. Average feature importance using SHAP values over the test folds. The amplitude of the SHAP value indicates the importance of the feature for the prediction (negative values means good outcome, positive values means bad outcome). The colours represent the value of the features, with red for high values (or true for binary) and blue for low values (or false for binary). The importance was calculated for Outcome 1 (A, B) and Outcome 2 (C, D) for the best performing models: LR (A, C) and SVM (B, D).

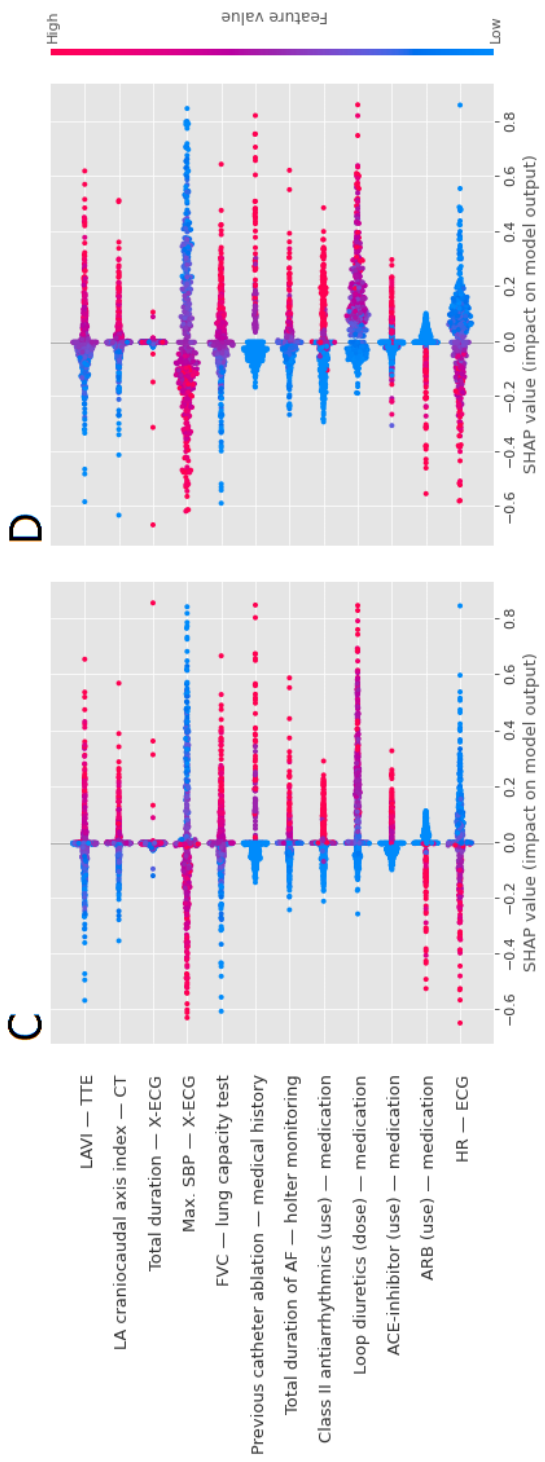

Figure 3. Continued

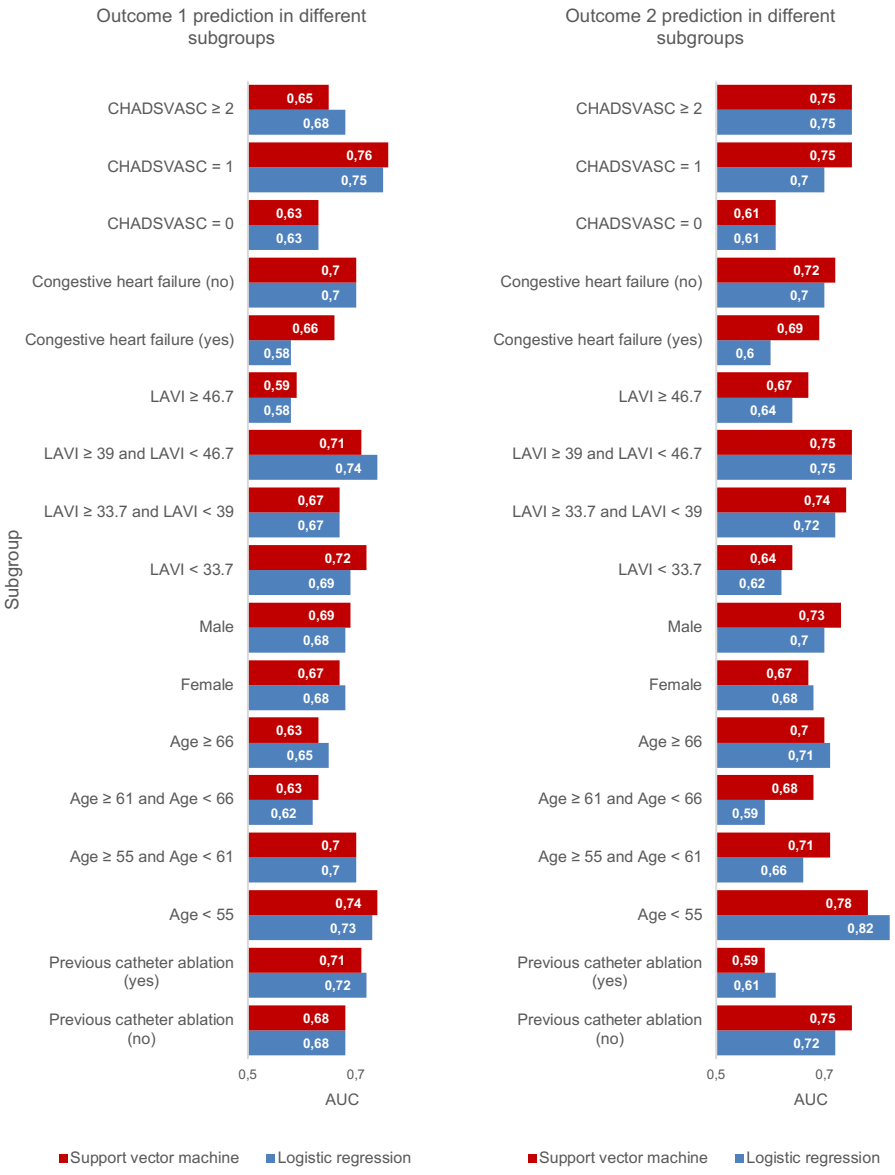

Figure 4. Subgroup analysis with pre-defined groups based on Outcome 1 and Outcome 2 best prediction models.

Prediction of AF Recurrence after Thoracoscopic Surgery

In line with risk scores and predictors for AF recurrence after CA for AF, clinical variables available before SA may predict which patients will benefit from SA. In this study, the use of all available baseline characteristics resulted in a moderate AUC to predict AF recurrence. However, an increased model performance was observed when using a selection of variables. A possible explanation is that input of a selection of key variables leads to less noise and redundancy. The key variables selected by LASSO to predict AF recurrence included LA size, which is a well-known predictor for AF recurrence after AF catheter ablation. Other included key variables were relatively uncommon as stand-alone predictors for AF recurrence. However, these may have been selected because they reflect patients' levels of frailty and comorbidities which may affect the risk of AF recurrence, or as a reflection (e.g., length) of well-known predictors (e.g., sex) that were not chosen. Surprisingly, patients with a low maximal SBP during exercise testing demonstrated to be at increased risk for AF recurrence. Possibly, this is because this group consists of the foremost advanced AF patients with a higher risk of AF recurrence, who are therefore more aggressively treated with antihypertensive or class II antiarrhythmic medication, or of patients with concomitant diastolic dysfunction. The selected key variables also explain why the model performs better in younger patients. As this patient group consists of patients with fewer comorbidities, it may represent a more homogeneous group with respect to the arrhythmogenic substrate for AF than older patients with multiple comorbidities.

AF Recurrence Definition and Measurement

Following current guidelines, AF recurrence was defined as any episode of atrial tachyarrhythmia lasting > 30 s beyond the three months blanking period (9). However, this definition is debatable, as one brief single episode does not carry the same symptom burden as episodes that last days to weeks (18). Our results did not show any difference in model performance when adjusting the definitions of AF recurrence. The models had a trend towards a higher AUC for Outcome 2 than for Outcome 1. A possible explanation is that recurrent AF may represent an advanced atrial substrate, or progressive disease, whereas recurrent atrial tachycardias may also result from technical failure of the procedure (i.e.,

reconnection across ablation lines) (19,20). However, due to the generally low burden of AF recurrence (21), repeat ablation was not performed in a large proportion of these patients and reconnection across ablation lines was not proven.

Additional Value of ML Techniques in the Prediction of AF Recurrence

It is expected that the application of ML techniques will improve future risk scores and prediction models. Our study shows a very moderate predictive value when using ML models including all available clinical variables as data input. However, using additional techniques, such as LASSO and SHAP, revealed some interesting findings that may improve prediction of AF recurrences after thoracoscopic AF surgery. Our findings underscore that ML tools, particularly those for selection and weighing of variables of interest, may contribute to improvement of prediction models and risk scores. This may be particularly relevant for large data sets with multiple variables wherein regular statistical methods show insufficient correlations.

Clinical Implications

Improved patient selection for SA could result in a higher success rate of the procedure. In patients with a predicted high risk of AF recurrence, it could be decided not to perform the procedure to prevent the associated complications. In addition, patient selection could identify patients at high risk for AF recurrence that could benefit from additional (continuous) monitoring, other specific follow-up management, and early re-intervention in case of (a)symptomatic AF recurrence. The selection of patients for SA is already based on a thorough preoperative screening based on the patient's medical history and baseline characteristics. Therefore, the included patients are already part of a highly selected population. This reduces the odds of improving patient selection with the available baseline variables, regardless of the use of ML techniques. As the AF field is evolving, future use of complex in-depth patient characteristics, procedural and mapping data, and improvements of the surgery technology, combined with different feature selection techniques, may further increase model performance.

Limitations

This study has some limitations. First, we only used data from a single center in our test and validation sets. Thereby, it is unknown how our models will perform in other comparable datasets. Furthermore, patients included in this analysis were patients who underwent SA. Patients that did not consent or were deemed unsuitable for the operation were therefore excluded from this analysis. This may impact on the generalizability of our findings. In addition, we did not perform a prospective validation of our models.

AF recurrence was monitored by repetitive ECGs and Holter monitoring as recommended by the guidelines (10). Patients were encouraged to obtain additional rhythm recording when symptomatic, but no continuous monitoring was performed. Therefore, asymptomatic recurrences of AF may have remained undetected. This could have been avoided by using loop recorders, which were not available for our population. However, the main goal of SA is to reduce AF-related symptoms in patients with advanced AF and thereby improve quality of life. Additionally, no specific indexes for adrenergic tone were available or included in this study. Finally, LASSO is, by definition, a linear regression with L1 regularization selecting features based on the linear correlation. As a result, the linear techniques might have been benefited when this feature selection was performed. The use of non-linear techniques (e.g., the feature importance of the RF) for feature selection, or even simpler techniques, might increase the accuracy of the ML techniques that can handle nonlinearities.

Conclusions

The proportion of risk of AF recurrence after SA embedded in baseline variables is modest. Advanced ML models predict recurrences of AF after SA best when using a selection of baseline characteristics, particularly in young patients.

References

1. Boersma LVA, Castella M, Boven Wv, Berruezo A, Yilmaz A, Nadal M, et al. Atrial Fibrillation Catheter Ablation Versus Surgical Ablation Treatment (FAST). Circulation. 2012;125(1):23-30.
2. Driessen AHG, Berger WR, Krul SPJ, van den Berg NWE, Neefs J, Piersma FR, et al. Ganglion Plexus Ablation in Advanced Atrial Fibrillation: The AFACT Study. J Am Coll Cardiol. 2016;68(11):1155-65.
3. Dretzke J, Chuchu N, Agarwal R, Herd C, Chua W, Fabritz L, et al. Predicting recurrent atrial fibrillation after catheter ablation: a systematic review of prognostic models. EP Europace. 2020;22(5):748-60.
4. Suzuki S, Yamashita T, Sakama T, Arita T, Yagi N, Otsuka T, et al. Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis. PLOS ONE. 2019;14(9):e0221911.
5. Rizwan A, Zoha A, Mabrouk IB, Sabbour HM, Al-Sumaiti AS, Alomainy A, Imran MA, Abbasi QH. A review on the state of the art in atrial fibrillation detection enabled by machine learning. IEEE reviews in biomedical engineering. 2020 Feb 27;14:219-39.
6. Lopes RR, van Mourik MS, Schaft EV, Ramos LA, Baan J, Jr., Vendrik J, et al. Value of machine learning in predicting TAVI outcomes. Neth Heart J. 2019;27(9):443-50.
7. Tang LYW, Ho K, Tam RC, Hawkins NM, Lim M, Andrade JG, editors. Predicting Catheter Ablation Outcomes with Pre-ablation Heart Rhythm Data: Less Is More. Machine Learning in Medical Imaging; 2020 2020//; Cham: Springer International Publishing.
8. Krul SP, Driessen AH, van Boven WJ, Linnenbank AC, Geuzebroek GS, Jackman WM, et al. Thoracoscopic video-assisted pulmonary vein antrum isolation, ganglionated plexus ablation, and periprocedural confirmation of ablation lesions: first results of a hybrid surgical-electrophysiological approach for atrial fibrillation. Circ Arrhythm Electrophysiol. 2011;4(3):262-70.
9. de Groot JR, Driessen AH, Van Boven WJ, Krul SP, Linnenbank AC, Jackman WM, et al. Epicardial confirmation of conduction block during thoracoscopic surgery for atrial fibrillation--a hybrid surgical-electrophysiological approach. Minim Invasive Ther Allied Technol. 2012;21(4):293-301.
10. Calkins H, Hindricks G, Cappato R, Kim YH, Saad EB, Aguinaga L, et al. 2017 HRS/EHRA/ECAS/APHRS/SOLAECE expert consensus statement on catheter and surgical ablation of atrial fibrillation. Heart rhythm. 2017;14(10):e275-e444.
11. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2011;28(1):112-8.
12. Pedregosa FV, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825--30.
13. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-88.
14. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018;300:70-9.

15. Njoku A, Kannabhiran M, Arora R, Reddy P, Gopinathannair R, Lakkireddy D, et al. Left atrial volume predicts atrial fibrillation recurrence after radiofrequency ablation: a meta-analysis. Europace. 2018;20(1):33-42.
16. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
17. Lundberg SL, S. A unified approach to interpreting model predictions. CoRR. 2017;abs/1705.07874.
18. Van Gelder IC, Healey JS, Crijns HJGM, Wang J, Hohnloser SH, Gold MR, et al. Duration of device-detected subclinical atrial fibrillation and occurrence of stroke in ASSERT. European Heart Journal. 2017;38(17):1339-44.
19. Bulava A, Mokracek A, Hanis J, Kurfirst V, Eisenberger M, Pesl L. Sequential hybrid procedure for persistent atrial fibrillation. Journal of the American Heart Association. 2015;4(3):e001754.
20. Kron J, Kasirajan V, Wood MA, Kowalski M, Han FT, Ellenbogen KA. Management of Recurrent Atrial Arrhythmias After Minimally Invasive Surgical Pulmonary Vein Isolation and Ganglionic Plexi Ablation for Atrial Fibrillation. Heart rhythm: the official journal of the Heart Rhythm Society. 2010;7(4):445-51.
21. Driessen AH, Berger WR, Chan Pin Yin DR, Piersma FR, Neefs J, van den Berg NW, et al. Electrophysiologically Guided Thoracoscopic Surgery for Advanced Atrial Fibrillation: 5-Year Follow-up. Journal of the American College of Cardiology. 2017;69(13):1753-4.

Supplementary Material

Table I. All variables and percentage of missing values.

Variable	Missing (%)	Variable	Missing (%)
AF duration (last episode) - *holter monitoring*	81	LIPV (height) - *CT*	12
AF duration (total) - *holter monitoring*	1	LIPV (width) - *CT*	69
Total ECV – *medical history*	24	LSPV (height) - *CT*	7
EHRA score – *symptom score*	33	LSPV (width) - *CT*	66
TV (diameter) - *TTE*	77	PV stenosis - *CT*	12
LAVI - *TTE*	37	RIPV (height) - *CT*	8
Creatinine - *blood sampling*	1	RIPV (width) - *CT*	69
CRP - *blood sampling*	19	RSPV (height) - *CT*	8
eGFR - *blood sampling*	4	RSPV (width) - *CT*	69
Hemoglobin - *blood sampling*	1	TV (diameter) - *CT*	10
INR- *blood sampling*	28	ACE-inhibitor (use) - *medication*	2
Potassium - *blood sampling*	0	ACE-inhibitor (dose) - *medication*	65
Leukocytes - *blood sampling*	1	ARB (use) - *medication*	2
Sodium - *blood sampling*	0	ARB (dose) - *medication*	70
NT-proBNP - *blood sampling*	7	Calcium antagonist (use) - *medication*	2
Thromobytes - *blood sampling*	1	Calcium antagonist (dose) - *medication*	78
Hs-troponine - *blood sampling*	42	Lipid lowering drugs (use) - *medication*	2
TSH - *blood sampling*	7	Lipid lowering drugs (dose) - *medication*	65
Urea - *blood sampling*	37	Class IA antiarrhythmics (use) - *medication*	3
Pulmonary FEV (absolute) - *lung capacity test*	14	Class IA antiarrhythmics (dose) - *medication*	84
Pulmonary FEV (relative) - *lung capacity test*	14	Class IC antiarrhythmics (use) - *medication*	2
Pulmonary FEV1VC (absolute) - *lung capacity test*	14	Class IC antiarrhythmics (dose) - *medication*	60
Pulmonary FEV1VC (relative) - *lung capacity test*	14	Class II antiarrhythmics (use) - *medication*	1
Pulmonary FVC (absolute) - *lung capacity test*	14	Class II antiarrhythmics (dose) - *medication*	44
Pulmonary FVC (relative) - *lung capacity test*	14	Class III antiarrhythmics (use) - *medication*	1
Arrhythmia - *X-ECG*	8	Class III antiarrhythmics (dose) - *medication*	53
Duration - *X-ECG*	7	Class IV antiarrhythmics (use) - *medication*	2

Ending - *X-ECG*	7	Class IV antiarrhythmics (dose) - *medication*	75
Ischemia - *X-ECG*	7	Other antiarrhythmics (use) - *medication*	2
Max. HR - *X-ECG*	6	Other antiarrhythmics (dose) - *medication*	74
Max. DBP - *X-ECG*	6	Loop diuretics (dose) - *medication*	76
Max. SBP - *X-ECG*	6	Loop diuretics (use) - *medication*	2
METS - *X-ECG*	45	Nitrates (use) - *medication*	3
Min. HR - *X-ECG*	7	Nitrates (dose) - *medication*	84
Min. DBP - *X-ECG*	7	OAC (use) - *medication*	0
Min. SBP - *X-ECG*	7	OAC (dose) - *medication*	79
WATT - *X-ECG*	57	Potassium diuretics (use) - *medication*	3
AVB - *ECG*	6	Potassium diuretics (dose) - *medication*	82
Axis - *ECG*	6	Thiazide diuretics (use) - *medication*	2
HR - *ECG*	4	Thiazide diuretics (dose) - *medication*	76
PR interval - *ECG*	24	Antiplatelet drug (use) - *medication*	3
QRS interval - *ECG*	4	Antiplatelet drug (dose) - *medication*	82
QT interval - *ECG*	5	Amiodarone - *medication*	13
QTc - *ECG*	5	Atenolol - *medication*	14
Rhythm - *ECG*	4	Bisoprolol - *medication*	13
Ventricular conduction - *ECG*	5	Carvedilol - *medication*	13
Aortic valve regurgitation - *TTE*	40	Digoxin - *medication*	14
Aortic valve stenosis - *TTE*	36	Diltiazem - *medication*	14
Mitral valve regurgitation - *TTE*	26	Disopyramide - *medication*	14
Mitral valve stenosis - *TTE*	42	Flecainide - *medication*	10
Pulmonary valve regurgitation - *TTE*	83	Quinidine - *medication*	14
Pulmonary valve stenosis - *TTE*	77	Metoprolol - *medication*	10
Tricuspid valve regurgitation - *TTE*	37	Nebivolol - *medication*	14
Tricuspid valve stenosis - *TTE*	52	Propafenone - *medication*	14
Type of failure - *holter monitoring*	0	Propranolol - *medication*	13
AF - *holter monitoring*	8	Sotalol - *medication*	11
Atrial flutter - *holter monitoring*	8	Verapamil - *medication*	14
Atrial tachycardia - *holter monitoring*	9	Age - *demographics*	0
AV block - *holter monitoring*	9	BMI – *physical examination*	0
Mean HR - *holter monitoring*	10	Height - *physical examination*	0

Max. HR - *holter monitoring*	8	HR - *physical examination*	1
Min. HR - *holter monitoring*	8	HR regular - *physical examination*	8
Flutter ablation - *medical history*	64	SBP - *physical examination*	2
Other cardiac procedure - *medical history*	9	Weight - *physical examination*	0
Cardiac surgery - *medical history*	0	DBP - *physical examination*	2
Catheter ablation, PVI - *medical history*	0	Age ≥ 65 - *demographics*	0
Catheter ablation, other - *medical history*	68	Age ≥ 75 - *demographics*	0
All PV - *Catheter ablation, PVI*	81	Alcohol - *intoxications*	55
Entry block - *Catheter ablation, PVI*	81	$CHADS_2$ - *risk score*	0
Exit block - *Catheter ablation, PVI*	81	CHA_2DS_2-VASc - *risk score*	0
Lesions - *Catheter ablation, PVI*	81	Hypercholesterolemia - *medical history*	20
CHF - *medical history*	9	Congestive heart failure - *medical history*	0
MI - *medical history*	0	Diabetes mellitus - *medical history*	0
Pacemaker - *medical history*	0	Drugs - *intoxications*	37
PCI - *medical history*	0	Family history of CVD - *medical history*	37
Surgical ablation - *medical history*	69	Female - *demographics*	0
Heart valve surgery - *medical history*	9	Hypertension - *medical history*	0
Aberrant PV - *CT*	4	Smoking - *intoxications*	22
LA anteroposterior axis index - *CT*	7	Stroke - *medical history*	0
LA craniocaudal axis index - *CT*	10	Vascular disease - *medical history*	0

Table II. Hyperparameters grid used for SVM.

Classifier	Kernel type	Penalty parameter *C*	Kernel coefficient γ	Degree of the Polynomial kernel	Class weight
SVM	Linear	[0.1, 1, 10, 100, 1000]	n.a.	n.a.	[None, Balanced]
	Radial basis function	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	n.a.	[None, Balanced]
	Polynomial	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	[3, 4, 5]	[None, Balanced]
	Sigmoid	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	n.a.	[None, Balanced]

Table III. Hyperparameters grid used for RF, GB, and NN.

Classifier	Parameter name	Parameter value
RF	Number of trees	[100, 500, 1000, 2000]
	Max features	[None, auto]
	Max depth	[None, 2, 3, 4]
	Min samples per split	[2, 4, 8]
	Min samples per leaf	[1, 2, 4]
	Class weight	[None, Balanced]
GB	Number of trees	[10, 50, 100, 200, 500]
	Max features	[None, auto]
	Max depth	[None, 2, 3, 4]
	Min samples per split	[2, 4, 8]
	Min samples per leaf	[1, 2, 4]
	Class weight	[None, Balanced]
NN	Activation	[relu]
	Hidden layer sizes	[5], [10], [50], [5, 5], [10, 10], [50, 50], [5, 5, 5], [10, 10, 10], [50, 50, 50]
	Alpha	[0.001, 0.0001]
	Solver	[adam]
	Learning rate	[adaptive]
	Initial learning rate	[0.1, 0.01, 0.001]

Table IV. Summarized patients characteristics for all included patients divided by outcome definition.

Grouped by outcome	Outcome 1 *(n=446)*		Outcome 2 (n=446)		Outcome 3 (n=446)		Outcome 4 (n=446)		Outcome 5 (n=446)	
	Success	Failure	Success	Failure	Success	Failure	Success	Failure	Success	Failure
n	*258*	*188*	*363*	*83*	*270*	*176*	*290*	*156*	*367*	*79*
Gender, n (%)										
Male	208 (80.6)	127 (67.6)	274 (75.5)	61 (73.5)	215 (79.6)	120 (68.2)	230 (79.3)	105 (67.3)	277 (75.5)	58 (73.4)
Female	50 (19.4)	61 (32.4)	89 (24.5)	22 (26.5)	55 (20.4)	56 (31.8)	60 (20.7)	51 (32.7)	90 (24.5)	21 (26.6)
BMI, mean (SD)	25.8 (7.2)	25.8 (7.8)	25.6 (7.8)	26.9 (5.8)	25.7 (7.2)	26.0 (7.8)	25.6 (7.6)	26.2 (7.3)	25.6 (7.7)	27.0 (5.9)
Age, mean (SD)	58.8 (8.7)	61.6 (8.4)	59.9 (8.6)	60.5 (9.1)	59.2 (8.7)	61.3 (8.5)	59.2 (8.9)	61.5 (8.2)	59.9 (8.6)	60.5 (9.3)
CHA_2DS_2-VASc, n(%)										
0	85 (32.9)	37 (19.7)	105 (28.9)	17 (20.5)	87 (32.2)	35 (19.9)	92 (31.7)	30 (19.2)	106 (28.9)	16 (20.3)
1	76 (29.5)	65 (34.6)	106 (29.2)	35 (42.2)	8 1 (30.0)	60 (34.1)	89 (30.7)	52 (33.3)	109 (29.7)	32 (40.5)
>=2	97 (37.6)	86 (45.7)	152 (41.9)	31 (37.3)	102 (37.8)	81 (46.0)	109 (37.6)	74 (47.4)	152 (41.4)	31 (39.2)
AF type										
Paroxysmal	130 (50.4)	50 (26.6)	154 (42.4)	26 (31.3)	137 (50.7)	43 (24.4)	141 (48.6)	39 (25.0)	157 (42.8)	23 (29.1)
Persistent	128 (49.6)	138 (73.4)	209 (57.6)	57 (68.7)	133 (49.3)	133 (75.6)	149 (51.4)	117 (75.0)	210 (57.2)	56 (70.9)

Table V. Average LR coefficients over folds. Positive values indicates bad outcome and negative values indicates good outcome.

Outcome 1		**Outcome 2**	
Variable	**Average**	**Variable**	**Average**
LAVI - *TTE*	0.011	LAVI - *TTE*	0.028
PR-interval - *ECG*	0.010	LA craniocaudal axis index - *CT*	0.020
RSPV (width) - *CT*	0.010	Duration - *X-ECG*	0.004
Tricuspid valve regurgitation - *TTE*	0.002	Max. SBP - *X-ECG*	-0.019
LA anteroposterior axis index - *CT*	0.011	FVC - *lung capacity test*	0.022
LA craniocaudal axis index - *CT*	0.036	Previous catheter ablation - *medical history*	0.372
Max. resistance - *X-ECG*	-0.003	AF duration (total) - *holter monitoring*	0.027
Max. SBP - *X-ECG*	-0.013	Class II antiarrhythmics (use) - *medication*	0.103
FEV1 - *lung capacity test*	-0.065	Loop diuretics (dose) - *medication*	0.036

FVC - *lung capacity test*	-0.103	ACE-inhibitor (use) - *medication*	0.054
Hs-troponine - *blood sampling*	-0.019	ARB (use) - *medication*	-0.140
Previous catheter ablation - *medical history*	0.439	HR - *ECG*	-0.019
Class II antiarrhythmics (use) - *medication*	-0.088		
Class III antiarrhythmics (dose) - *medication*	-0.001		
ACE-inhibitor (use) - *medication*	0.028		
Age - *demographics*	0.024		
Height - *physical examination*	-0.679		
Type of AF - *medical history*	0.621		

7

Machine learning-based prediction of insufficient contrast enhancement in coronary computed tomography angiography

Lopes RR, van den Boogert TP, Lobe NH, Verwest TA, Henriques JP, Marquering HA, Planken RN.

European radiology. 2022 Oct;32(10):7136-45.

DOI: 10.1007/s00330-022-08901-5

Abstract

Patient-tailored contrast delivery protocols strongly reduce the total iodine load and in general improve image quality in CT coronary angiography (CTCA). We aim to use machine learning to predict cases with insufficient contrast enhancement and to identify parameters with the highest predictive value.

Machine learning models were developed using data from 1,447 CTs. Were included patient features, imaging settings and, test bolus features. The models were trained to predict CTCA images with a mean attenuation value in the ascending aorta below 400 HU. The accuracy was assessed by the area under the receiver operating characteristic (AUROC) and precision-recall curves (AUPRC). Shapley Additive exPlanations, was used to assess the impact of features on the prediction of insufficient contrast enhancement.

A total of 399 out of 1,447 scans revealed attenuation values in the ascending aorta below 400 HU. The best model trained using only patient features and CT settings achieved an AUROC of 0.78 (95% CI: 0.73–0.83) and AUPRC of 0.65 (95% CI: 0.58–0.71). With the inclusion of the test bolus features, it achieved an AUROC of 0.84 (95% CI: 0.81–0.87), an AUPRC of 0.71 (95% CI: 0.66–0.76), and a sensitivity of 0.66 and specificity of 0.88. The test bolus' peak height was the feature that impacted low attenuation prediction most.

Prediction of insufficient contrast enhancement in CT coronary angiography scans can be achieved using machine learning models. Our experiments suggest that test bolus features are strongly predictive of low attenuation values and can be used to further improve patient-specific contrast delivery protocols.

Introduction

Computed tomographic coronary angiography (CTCA) is a non-invasive imaging technique used for the anatomical assessment of coronary artery disease [1–6]. Iodine containing contrast material (CM) is used to enhance luminal attenuation to enable assessment of the coronary artery lumen, vessel wall, and the surrounding structures [7]. Adjustments in CM delivery protocols change the attenuation coefficient of the blood pool. A commonly used strategy to adjust CM delivery is to regulate the iodine delivery rate (IDR = amount of iodine injected per second [g I/s] = concentration of CM × flow rate in ml/s) [7]. Besides CM delivery, coronary lumen attenuation also depends on patient features like body weight and length as well as CT scanner settings and the tube voltage (kV) in particular [7]. Other parameters, such as the peak height and time to peak of a test bolus, are also associated with attenuations but are commonly not considered in current CM protocols [8, 9]. A better understanding of the interrelation between these parameters and luminal attenuation is valuable for further improvements in patient-specific contrast delivery protocols. Reducing the iodine load is important to lower the risk for renal function impairment, reduce environmental pollution, and lower overall costs. However, inappropriate correction in contrast administration may result in insufficient coronary lumen attenuation and this is not tolerable.

For accurate assessment of coronary artery disease on CTCA, intra-arterial attenuation values higher than 350 HU are recommended [10–15]. In previous studies, the introduction of patient-tailored CM protocols, adjusting the IDR for body weight and kV, resulted in more constant coronary artery attenuation values and a favorable reduction in total iodine load [8, 10–12]. However, in some cases, CM delivery resulted in low coronary attenuation values, thereby jeopardizing the diagnostic value of CTCA [8].

We hypothesized that machine learning (ML) can help to predict cases with insufficient contrast attenuation in CTCA. This will allow for CM delivery and CT scanner settings to enhance coronary attenuation and improve diagnostic value. Additionally, we investigated the added value of using the test bolus features for the prediction. To this end, we also analyzed the impact of the features on predicting insufficient attenuation.

Materials and Methods

Study design and population

This retrospective study was performed following the principles of the Declaration of Helsinki and the local Institutional Review Board approved this study. The Ethics Committee approved this research with a waiver. All consecutive patients above 18 years old who underwent CTCA between September 2017 and September 2020 were included in the study. CT scans were excluded if the acquisition protocol deviated from the standard CTCA protocol (e.g., TAVI or cardiac function) or if the test bolus enhancement curves were not stored in the hospital's picture archiving and communication system.

CTCA acquisition protocol

The imaging protocol has been described before [8]. In summary, all images were obtained using a third-generation dual-source 192 detector row CT scanner (Somatom Force, Siemens Healthcare). Sublingual nitro-glycerine spray was administered before the CTCA acquisition and beta-blockers were administered on indication (heart rate > 65 per min). The time between the start of contrast medium injection and the time to peak contrast enhancement in the ascending aorta was determined using a test bolus injection with a fixed contrast bolus of 10 ml undiluted contrast medium (Ultravist 300: iopromide 300 mg I/ml, Bayer AG or Xenetix 350: iobitridol 350 mg I/ml, Guerbet OptiVantage DH) a fixed scan-dealy of 8 s and a fixed kV value of 100 kV. For timing the CTCA acquisition, the scan delay was determined by the time to peak and an additional 4 s for coronary artery filling. For the CTCA scans, automatic tube voltage selection (CARE kV, Siemens Healthcare, Erlangen, Germany) was applied in all patients with kV categories ranging from 70 to 120kV with increments of 10kV. All CTCA scans were visually evaluated by the attending CT technician. CT scanner acquisition parameters were: detector collimation 2 × 96 × 0.6 mm, slice acquisition 2 × 192 × 0.6 mm using a z-flying focal spot, gantry rotation time of 250 ms, temporal resolution of 66 ms, 70–120 kV tube voltage (CARE kV), and 180-600 μA tube current. High-pitch spiral scanning was performed in diastole in patients with a regular heart rate < 70/min. A prospective ECG-gated sequential scan (step and shoot) was performed in diastole for patients with irregular heart rate < 70/min or heart rates ranging between 70 and 80/min. For

patients with irregular heart rates of > 80/min, a sequential scan was performed in systole. Padding in an adaptive prospective sequential scan mode for high and irregular heart rates was used to enable reconstruction of more cardiac phases. Images were reconstructed with a slice thickness of 0.6 mm and an increment of 0.4 mm using iterative reconstruction factor 2 (ADMIRE, Siemens Healthcare).

Contrast delivery protocol

Iodinated contrast medium (300 or 350 mg I/ml) was administered via a dual-head contrast delivery injector (Guerbet OptiVantage DH) equipped with a high-pressure resistant extension tube and injected in the right antecubital vein. A test bolus of 10 ml contrast medium was injected at 6 ml/s or 6.5 ml/s, followed by a 40 ml saline chaser also injected at 6 or 6.5 ml/s. The bolus of (un)diluted contrast material for high-pitch spiral CTCA scans was 50 ml and the bolus for prospective sequential step-and-shoot scans was 65 ml. The larger contrast bolus volume in prospective sequential step-and-shoot scans was applied to compensate for the longer acquisition time. The contrast bolus was injected at an injection rate of 6 ml/s or 6.5ml/s. All contrast injections were followed by a saline chaser of 40 ml (6 or 6.5 ml/s). The IDR was adjusted for body weight and kV settings, as presented in a previous study [8]. The kV settings for CTCA acquisition, as selected by CARE kV, were used together with body weight to provide a patient-specific IDR (1–2.3 g I/s). To reach the required IDR, the CM was diluted with saline via the dual-head contrast delivery injector, of which one was filled with undiluted contrast material and the other with saline solution. The two fluids were blended in the high-pressure resistant extension tube, after which it was injected in the right antecubital vein.

Data extraction

Data was retrieved automatically from DICOM headers and electronic patient records. Collected patient features included sex, age, average heart rate, body weight, and body height. Also, kV settings (tube voltage), iodine delivery rate (IDR), total iodine load, and contrast dose concentration were collected. Furthermore, we extracted test bolus features, such as the peak height of the test bolus attenuation curve (peak height in HU), the time to peak of the contrast curve (time-to-peak in seconds), and the time to the start of the contrast curve (the time-to-start-curve in seconds) from the bolus tracking curves (DynEva, Siemens

Healthcare) as illustrated in Figure 1. An association between the height of the test bolus and coronary attenuation has been reported in previous studies. Therefore, we considered the test bolus to contain important information and included this in the model [8, 9]. Regarding the time to start and time to peak, the default delay of 8 s was ignored in the analysis once the values used were obtained from the bolus tracking curves.

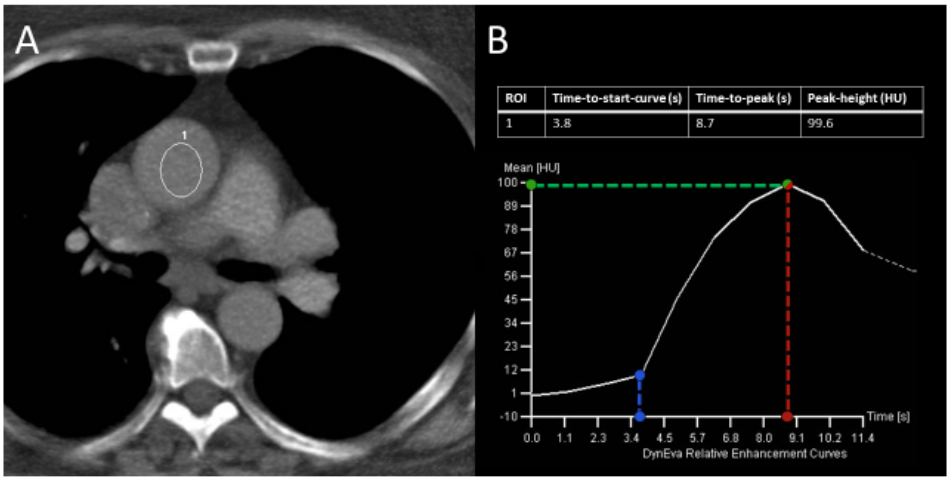

Figure 1. Dynamic bolus tracking of the test-bolus scan example. An ROI is used to measure the attenuation at the level of the ascending aorta below the level of the carina (A, 1). The curve (B) represents the measured values over time. Time-to-start-curve (blue dashed line) in seconds, time-to-peak (red dashed line) in seconds, and peak height (green dashed line) in HU where t = 0 corresponds to 8 seconds after contrast media injection.

For the assessment of luminal attenuation, we used an in-house developed tool to automatically detect the ascending aorta. Correspondingly, a region of interest with a radius of approximately 70% of the aorta radius was fitted to calculate the average attenuation in the ascending aorta and exclude possible edges and calcifications in the vessel wall. In cases in which the tool did not detect the ascending aorta, the location was selected manually. The attenuation value in the ascending aorta was used as a proxy of the attenuation in the coronary arteries. It should be noted that the attenuation in the ascending aorta is slightly higher than that in the coronary arteries. Previous research has shown that there is a strong

association between attenuation in the ascending aorta and coronary arteries with a mean decay of 25 HU expected from the ascending aorta to the proximal coronary arteries and of 50 HU to the distal coronary arteries [8]. Therefore, the cutoff value for adequate attenuation in the ascending aorta for this study was 400 HU.

Model development

The models were trained to predict insufficient luminal attenuation in the ascending aorta. Insufficient attenuation was defined as an average attenuation lower than 400 HU within the region of interest. ML techniques were used to deal with both linear and nonlinear interactions between the included features. These techniques included the following: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB), support vector machines (SVM), and neural networks (NN). To assess the added value of extracting information from the test bolus, we performed two experiments: only patient features with CT settings and, additionally, also including the test bolus features. Both experiments followed the same methods and only differed in the features included.

We used stratified 10-fold cross-validation (CV) for the development and evaluation of the ML models. In some cases where CTs from the same patient were split into training and test set, the CTs were removed from the training set to avoid patient data leakage. The training set was also used to find the optimal hyper-parameters using a grid search with another 5-fold CV. For model selection after the hyper-parameter optimization, the models with the largest average area under precision (positive predictive value) and recall (sensitivity) curves were selected. The testing folds were not used during the training steps.

To deal with the missing values, we used MissForest, an iterative technique based on random forests [16] for imputation. The imputation model was created with the training data only to avoid data leakage. As a requirement for some of the ML techniques, the continuous features were standardized by removing their mean and scaling to unit variance.

Detailed information about the selected classifiers and hyper-parameters used for optimization is available in the Supplementary Material Tables I and II. The

analysis was performed with Python (Python Software Foundation, version 3.6, www.python.org) using the scikit-learn [17] and XGBoost [18] packages.

Model evaluation

The area under the receiver operating characteristic (AUROC) and precision-recall curves (AUPRC) were used to evaluate the models. As a 10-fold CV was applied, we computed the averages and 95% confidence interval (CI) for each model. The Wilcoxon signed-rank test was performed to assess whether the difference in AUROC and AUPRC between the prediction models with and without using the test bolus features are statistically significant (p-value < 0.05).

Model interpretation

For the visualization of the importance of included features in the prediction analysis, the Shapley Additive exPlanations (SHAP) framework was used [19]. For each of the features, the feature importance (SHAP value) was calculated by making predictions excluding that feature. This value describes how it affects the prediction probability. The larger the SHAP value, the more it affects the prediction. Additionally, the values can be either positive, for low attenuations, or negative, for regular attenuations.

The SHAP values were computed for the entire population. In addition, for a better understanding of the effect of the features per tube voltage, we also computed the SHAP values per tube voltage group (70 – 120 kV).

Results

A total of 1,447 scans from 1,364 patients were included in the analysis. Of these scans, 399 (27%) were considered to have insufficient attenuation. Figure 2 displays an example of CTCAs with insufficient (227 HU), accurate (433 HU), and high (595 HU) attenuation. The tool for automated ascending aorta detection failed in less than 1% of the cases. Baseline and descriptive features are shown in Table 1 and average attenuation values per tube voltage are shown in Table 2. The relationship between patient weight and mean attenuation per kV group is shown in Supplementary Material Figure I. Despite the already-applied correction for kV and body weight in our acquisition protocol, there was considerable variation between patients.

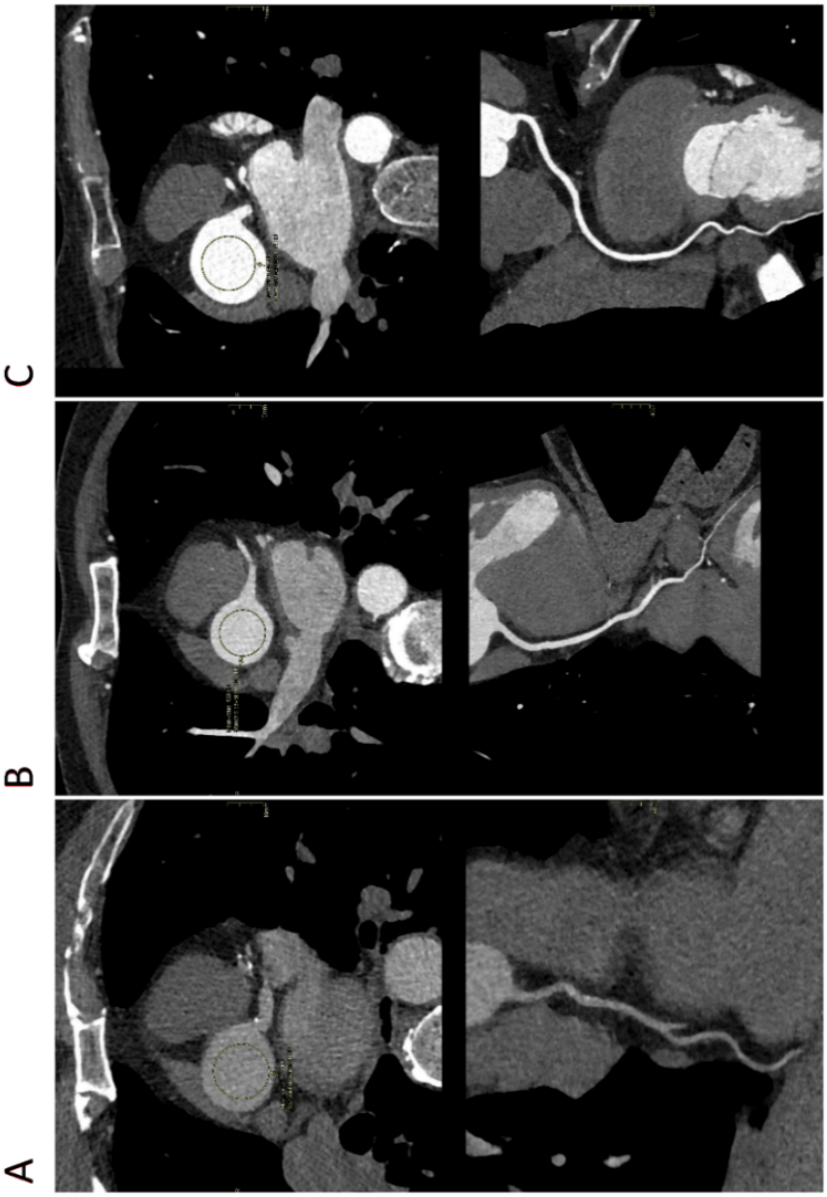

Figure 2. Examples of CTCA scan (axial slice through the ascending aorta above and curved multiplanar reconstruction of the right coronary artery below of patients with low (A), accurate (B), and high (C)

Table 1. Descriptive statistics of the study group, mean ± SD or N (%).

		Missing (n)	Low attenuation (n=399)	Regular attenuation (n=1048)
Age (yrs)		374	55.7 ± 11.8	52.8 ± 12.2
Sex (female)		374	494 (62%)	101 (37%)*
Height (cm)		568	171 ± 10	177 ± 10*
Weight (kg)		556	77 ± 14	86 ± 19*
Average heart rate (bpm)		0	61.9 ± 10.2	61.6 ± 11.2
Iodine delivery rate (g I/s)		0	1.5 ± 0.3	1.5 ± 0.3*
Tube voltage (kV)	**70**	0	357 (34%)	111 (28%)*
	80		426 (41%)	110 (28%)
	90		214 (20%)	75 (19%)
	100		29 (3%)	39 (10%)
	110		18 (2%)	23 (6%)
	120		4 (0%)	41 (10%)
Total iodine load (g)		0	16.0 ± 2.8	15 ± 2.7
Peak height - test bolus (HU)		0	127 ± 40	95 ± 35*
Time to peak - test bolus (s)		0	8.5 ± 2.7	9.6 ± 3.4*
Time to start - test bolus (s)		2	3.4 ± 1.9	4.3 ± 2.4*

*p<0.001, Two-sample T-test or Chi-square, as appropriate.

Table 2. Average and standard deviation of the attenuations per tube voltage group.

	Low attenuation		Regular attenuation	
Tube voltage (kV)	**n**	**Attenuation (HU)**	**n**	**Attenuation (HU)**
70	111	350 ± 37	357	522 ± 41
80	110	342 ± 35	426	499 ± 37
90	75	348 ± 34	214	506 ± 36
100	39	347 ± 35	29	484 ± 36
110	23	340 ± 27	18	473 ± 33
120	41	312 ± 36	4	431 ± 44

The AUROC and AUPRC together with corresponding 95% CI for all experiments are shown in Table 3. The models with the highest average accuracies were trained with RF and had an AUPRC of 0.71 (95% CI: 0.66–0.76) and AUROC of 0.84 (95% CI: 0.81–0.87), with a sensitivity of 0.66 and

specificity of 0.88 (Figure 3). Notably, these models included the test bolus features. Regarding the models without the test bolus features, the best performing model was also achieved by the RF model. In comparison, this model had an AUROC of 0.78 (95% CI: 0.73–0.83) and AUPRC of 0.65 (95% CI: 0.58–0.71). The differences between the AUROC and AUPRC values for the various models using the test bolus features were not statistically significant. The AUROC difference of the prediction models between using and not using the test bolus features was statistically significant (p-value = 0.027). The difference in AUPRC between the two models was not statistically significant (p-value = 0.23).

Table 3. Evaluation of the low attenuation detection models with 95% confidence interval. AUPRC = area under the precision-recall curve, AUROC = area under the receiver operating characteristic curve.

	Including patients features CT settings and test bolus features		Including patients features and CT settings	
Model/Metric	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**
Logistic regression	0.70 (0.63-0.76)	0.83 (0.79-0.87)	0.62 (0.55-0.68)	0.77 (0.72-0.82)
Random forest	0.71 (0.66-0.76)	0.84 (0.81-0.87)	0.65 (0.58-0.71)	0.78 (0.73-0.83)
XGBoost	0.70 (0.66-0.75)	0.83 (0.80-0.87)	0.64 (0.59-0.69)	0.78 (0.74-0.82)
Support vector machines	0.67 (0.62-0.73)	0.82 (0.78-0.86)	0.56 (0.50-0.63)	0.75 (0.70-0.80)
Neural networks	0.69 (0.63-0.74)	0.82 (0.79-0.86)	0.61 (0.55-0.68)	0.76 (0.71-0.80)

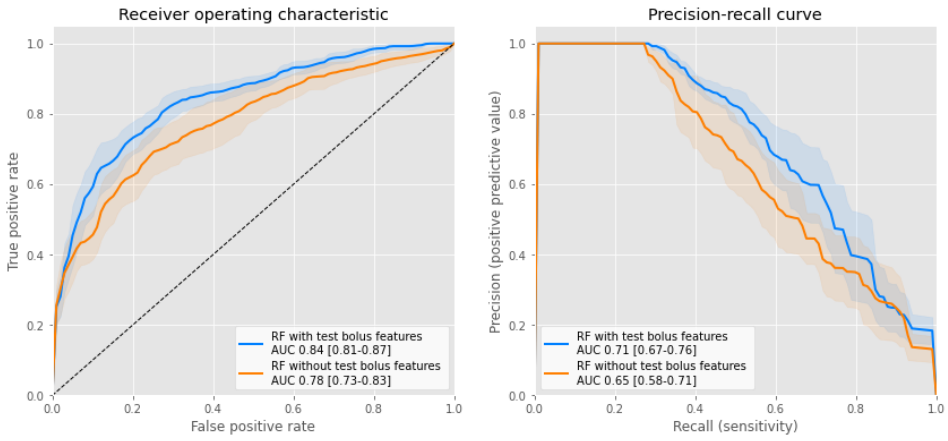

Figure 3. Average receiver operating characteristic (left) and precision-recall curves (right) with 95% confidence interval for models trained with patient features, CT settings,

and test bolus features and with only patient features and CT settings. RF = *random forest, AUC = area under the curve.*

The SHAP summary plot is presented in Figure 4, showing only the features related to the CM protocol. As might be expected, it shows that high tube voltages are strongly associated with low attenuations. Furthermore, higher body weights and lower IDR also result in higher chances of insufficient attenuation. As the contrast delivery protocol, used in this study, adjusted the IDR for kV settings and body weight, the effect of these features was evaluated in each tube voltage group (Figure 5). The kV categories of 70, 80, and 90 kV are associated with intended attenuation values (with a negative SHAP value) and 100, 110, and 120 kV with lower attenuation values (with a positive SHAP value).

Figure 6 shows the SHAP values of all features used in the model. Of all these features, the peak height of the test bolus contrast curve is the most impactful feature (with low peak height associated with low attenuation) followed by body height (high body height values associated with low attenuation). Regarding the protocol features, the tube voltage is the third most important.

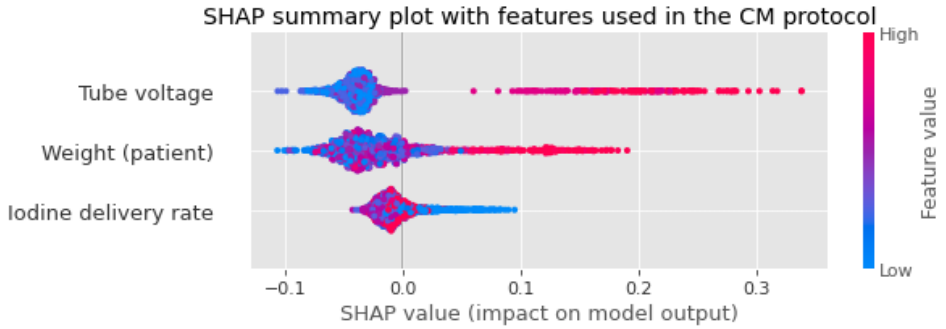

Figure 4. Importance of the CM protocol features (average on the test folds) using SHAP values. The amplitude of the SHAP value indicates the feature importance for the prediction (positive values mean low attenuation). The colors represent the values of the features, with red for high values and blue for low values. *CM = Contrast material.*

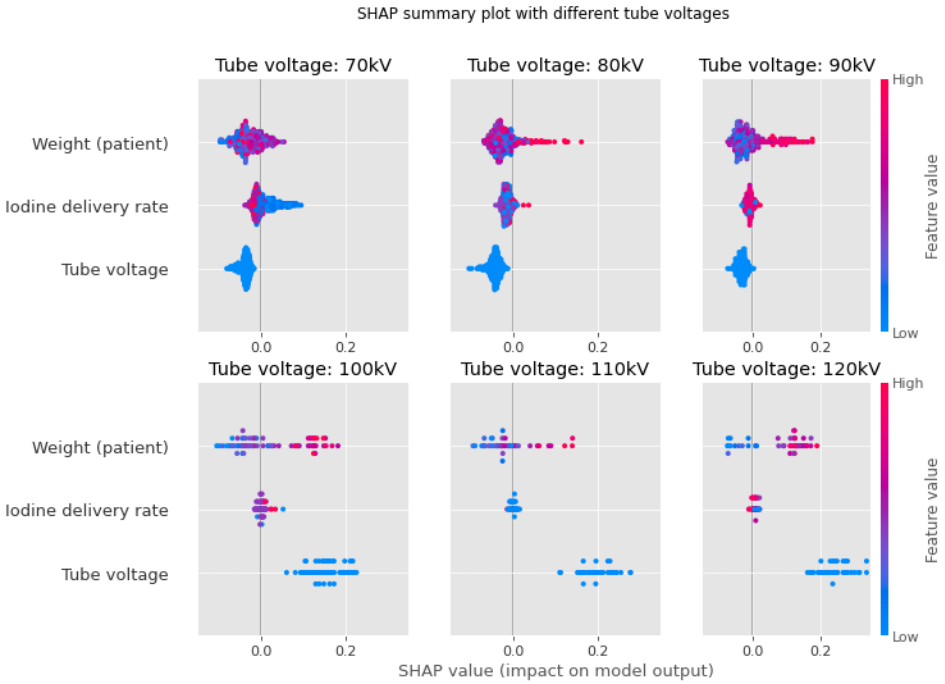

Figure 5. Importance of the CM protocol features (average on the test folds) using SHAP values divided by tube voltage group. The amplitude of the SHAP value indicates the feature importance for the prediction (positive values mean low attenuation). The colors represent the values of the features, with red for high values and blue for low values. Note that there is only one color for the tube voltage since there is only one tube voltage per group. *CM = Contrast material*.

Discussion

In this study, we have shown that ML models are accurate in predicting low attenuation scans. Moreover, in the setting of a patient-specific contrast delivery protocol adjusting the IDR for kV setting and body weight, the peak height of the test bolus curves is the most impacting feature for the model. Including the test bolus features, the prediction accuracy of the models increased, compared to models using only patient features and CT settings. This highlights the association of the test bolus features, specifically the peak height of the test bolus attenuation curve with luminal attenuation and this should be considered when further refining contrast protocols.

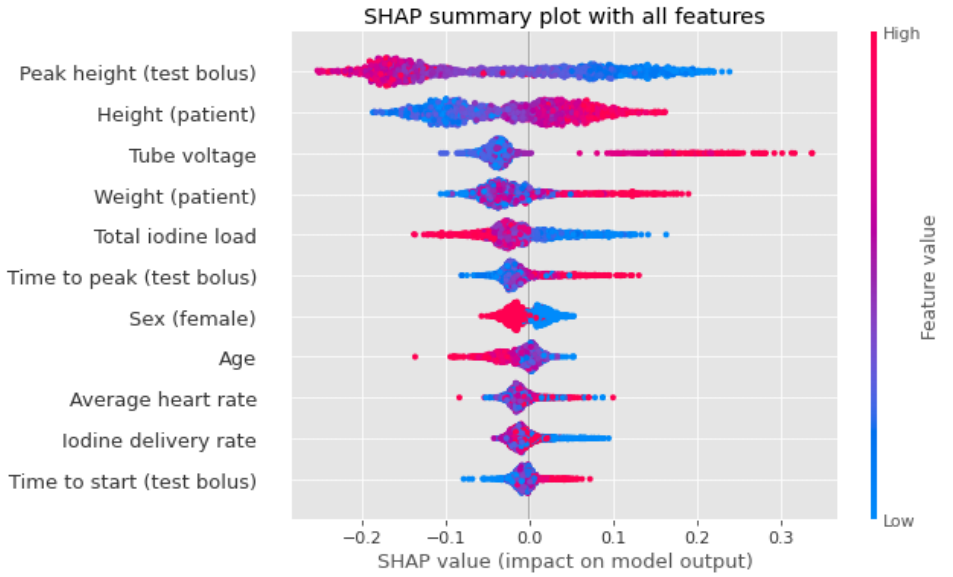

Figure 6. Importance of all features included in the model (average on the test folds) using SHAP values. The amplitude of the SHAP value indicates the feature importance for the prediction (positive values mean low attenuation). The colors represent the values of the features, with red for high values and blue for low values. *CM = Contrast material.*

In our population, attenuation was inversely associated with kV, despite IDR adjustment for kV settings. These results suggest that it is worth adjusting the IDR even more for kV in our clinical protocol. However, such extra IDR adaptation for kV cannot account for the interplay of other settings on IDR and image quality as indicated by the results showing that multiple parameters influence image quality, most likely in a non-linear fashion. In a previous study, we showed that the patient-tailored contrast delivery protocol contributed to reduced variation in contrast attenuation in the coronary arteries. However, despite the correction for kV and body weight, there remained considerable variation between patients, and coronary attenuation was not sufficiently high in all patients to assure accurate radiologic assessment. Therefore, our study suggests that a straightforward correction for kV settings and body weight underestimates the complexity of the scanning parameters, which do not take the interaction of other parameters with the IDR into account.

Although the peak height of the test bolus was already found significantly associated with image quality in other studies, in this study, we aimed at the

application of AI to predict too low coronary artery attenuation in a clinical setting with a contrast protocol adjusted for body weight and kV. The strong association of the test bolus and optimal enhancement in the ascending aorta on the CCTA is not surprising because of its similar signal. However, the test bolus is a small volume of contrast. A longer bolus results in accumulation and therewith a higher plateau of attenuation. The filling time will result in this plateau feature. The form of this upslope and plateau may vary, most likely concerning time to peak and peak value.

Model performance

We evaluated five different ML techniques and the differences between the accuracy of these models were not statistically significant. All evaluated ML techniques used in this study seem to be able to identify insufficient contrast cases, including LR, which only takes linear relationships between features and outcome into consideration. Using as reference the RF model, with a sensitivity of 0.66 and specificity of 0.88, we can identify 263 (from 399) CTs with a relatively small number of false positives (125). Additionally, the prediction probability threshold could be adjusted to have higher sensitivity at a cost of lowering the specificity.

Comparison with previous studies

Multiple studies aim to use ML to improve the CT acquisition process and image quality [20]. Also, some studies aimed on developing patient-tailored CM protocols using the test bolus features in 100–120 kV scans, not covering the currently available kV range 70–120 kV [23, 24].

Besides the use of test bolus, some studies use tailored CM protocols with (automatic) bolus tracking. Martin et al. [25] evaluated the feasibility of a vendor's software using a tube voltage-tailored CM application, which still resulted in more than 25% of the CTAs with attenuations in the ascending aorta below 400 HU. In another study with bolus tracking, Yin et al. [26] evaluated protocols tailored for BMI or BSA, and, either way, cases of insufficient attenuation in the aorta occurred. The use of AI, as presented in the current study, could potentially improve different protocols by automatically detecting cases with insufficient attenuation when using a test bolus protocol.

Limitations

This study was performed with a relatively large cohort; however, due to the retrospective nature of the study design, some patient-specific features were incomplete. Furthermore, this is a single-center study and the CTCAs were acquired with a specific protocol, making the ML models not generally suitable for different protocols without re-training them with additional data. Also, the selected cutoff value, 400 HU for the ascending aorta is arbitrary. However, it should be noted that this value is not the only marker of high-quality coronary CTA. Moreover, the quality was addressed by the (objective) attenuation assessment whereas the quality could also have been addressed by the (subjective) radiologist's rating. However, it should be noted that this value is not the only marker of high-quality coronary CTA.

Regarding the ML techniques used in this study, all models tested achieved similar accuracies. It might be explained by the limited number of features considered for this analysis. The addition of more features, such as information extracted from the test scan or engineered features, would add additional value that could be exploited by the techniques that can handle a large number of features and nonlinearities. Although attenuation is an important topic regarding image quality of CCTAs, it does not cover image quality completely. Noise, artifacts, or qualitative quality assessments were not considered in this study.

The model can accurately predict low attenuation retrospectively in a large population. Therefore, the current study should be conceived as a proof-of-concept study to predict low attenuation. The effectiveness of the proposed prediction model needs to be addressed in a subsequent prospective study. Also, the extent to which the IDR should be adjusted was beyond the scope of this study. Regression models to estimate the attenuation itself, instead of a binary classification, may be a solution. With a correct estimation of the attenuation, the IDR could be adjusted such that the predicted attenuation is close to acquired luminal attenuation that should be close to the desired value. This approach will be explored in a further study.

Although the peak height of the test bolus was already found significantly associated with image quality in other studies [8, 9], in this study, we aimed at

the application of AI to predict too low coronary artery attenuation in a clinical setting with a contrast protocol adjusted for body weight and kV. The strong association of the test bolus and optimal enhancement in the ascending aorta on the CCTA is not surprising because of its similar signal. However, the test bolus is a small volume of contrast. A longer bolus results in accumulation and therewith a higher plateau of attenuation. The filling time will result in this plateau feature. The form of this upslope and plateau may vary, most likely concerning time to peak and peak value.

An infrequent but important factor for inadequate contrast media arrival is dynamic venous compression in the thoracic outlet region. Although technicians are trained in patient positioning to avoid venous compression for optimal contrast dynamics, dynamic venous compression can not be ruled out entirely and this might have also contributed to low attenuation in some patients, which is not accounted for in the current analysis.

Clinical implications

Current fixed CM delivery protocols may be too simple for adequate contrast enhancement in CTCA. An important step in improving patient-tailored contrast delivery protocols is to understand why and when current approaches fail. Predicting when the protocol is potentially failing is the first step to develop more robust protocols. With the models developed in this study, insufficient attenuation can accurately be predicted and adjustments (such as increasing the IDR) can be performed to avoid too low attenuation. This study also shows the potential value of the information that can be extracted from the test bolus which can be incorporated in more advanced and robust protocols.

Conclusion

We demonstrate that ML is accurate in the prediction of CCTA with insufficient attenuation on our local imaging protocol. We have shown that, in a protocol already adjusting for kV and body weight, the most impacting feature for the ML model is the peak height of the test bolus curve. Our findings support the development of more refined and more robust patient-tailored contrast delivery protocols with the inclusion of test bolus features. Also, it should be noted that

the approach is general and could be applied to a wide range of scanning protocols.

References

1. Mollet NR, Cademartiri F, van Mieghem CAG, et al (2005) High-resolution spiral computed tomography coronary angiography in patients referred for diagnostic conventional coronary angiography. Circulation 112:2318–2323
2. Achenbach S, Giesler T, Ropers D, et al (2001) Detection of coronary artery stenoses by contrast-enhanced, retrospectively electrocardiographically-gated, multislice spiral computed tomography. Circulation 103:2535–2538
3. Meijboom WB, Meijs MFL, Schuijf JD, et al (2008) Diagnostic accuracy of 64-slice computed tomography coronary angiography: a prospective, multicenter, multivendor study. J Am Coll Cardiol 52:2135–2144
4. Nieman K, Oudkerk M, Rensing BJ, et al (2001) Coronary angiography with multi-slice computed tomography. Lancet 357:599–603
5. Marano R, Rovere G, Savino G, et al (2020) CCTA in the diagnosis of coronary artery disease. Radiol Med 125:1102–1113. https://doi.org/10.1007/s11547-020-01283-y
6. W. SP, Hironori H, Scot G, et al (2021) Coronary Computed Tomographic Angiography for Complete Assessment of Coronary Artery Disease. J Am Coll Cardiol 78:713–736. https://doi.org/10.1016/j.jacc.2021.06.019
7. Mihl C, Maas M, Turek J, et al (2017) Contrast media administration in coronary computed tomography angiography–a systematic review. In: RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren. © Georg Thieme Verlag KG, pp 312–325
8. van den Boogert TPW, Lopes RR, Lobe NHJ, et al (2021) Patient-tailored Contrast Delivery Protocols for Computed Tomography Coronary Angiography: Lower Contrast Dose and Better Image Quality. J Thorac Imaging. https://doi.org/10.1097/RTI.0000000000000593
9. Tan SK, Ng KH, Yeong CH, et al (2019) Personalized administration of contrast medium with high delivery rate in low tube voltage coronary computed tomography angiography. Quant Imaging Med Surg 9:552
10. Isogai T, Jinzaki M, Tanami Y, et al (2011) Body weight-tailored contrast material injection protocol for 64-detector row computed tomography coronary angiography. Jpn J Radiol 29:33–38
11. Fei X, Du X, Yang Q, et al (2008) 64-MDCT coronary angiography: Phantom study of effects of vascular attenuation on detection of coronary stenosis. Am J Roentgenol. https://doi.org/10.1016/j.apjtb.2017.07.017
12. Nakaura T, Awai K, Yauaga Y, et al (2008) Contrast injection protocols for coronary computed tomography angiography using a 64-detector scanner: comparison between patient weight-adjusted-and fixed iodine-dose protocols. Invest Radiol 43:512–519
13. Yamamuro M, Tadamura E, Kanao S, et al (2007) Coronary angiography by 64-detector row computed tomography using low dose of contrast material with saline chaser: influence of total injection volume on vessel attenuation. J Comput Assist Tomogr 31:272–280
14. Cademartiri F, Maffei E, Palumbo AA, et al (2008) Influence of intra-coronary enhancement on diagnostic accuracy with 64-slice CT coronary angiography. Eur Radiol. https://doi.org/10.1007/s00330-007-0773-0

15. Bae KT, Tran HQ, Heiken JP (2004) Uniform vascular contrast enhancement and reduced contrast medium volume achieved by using exponentially decelerated contrast material injection method. Radiology 231:732–736
16. Stekhoven DJ, Bühlmann P (2012) MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112–118
17. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830
18. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 785–794
19. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. pp 4768–4777
20. Eberhard M, Alkadhi H (2020) Machine learning and deep neural networks: applications in patient and scan preparation, contrast medium, and radiation dose optimization. J Thorac Imaging 35:S17–S20
21. Wang Y, Yu M, Wang M, et al (2019) Application of Artificial Intelligence–based Image Optimization for Computed Tomography Angiography of the Aorta With Low Tube Voltage and Reduced Contrast Medium Volume. J Thorac Imaging 34:393–399
22. Ghoshhajra BB, Engel L-C, Károlyi M, et al (2013) Cardiac computed tomography angiography with automatic tube potential selection: effects on radiation dose and image quality. J Thorac Imaging 28:40–48
23. Zhu X, Zhu Y, Liu W, et al (2016) Improved image-quality consistency in coronary CT angiography using a test-bolus-based individually tailored contrast medium injection protocol. Clin Radiol 71:1113–1119
24. Kidoh M, Nakaura T, Nakamura S, et al (2014) Novel contrast-injection protocol for coronary computed tomographic angiography: contrast-injection protocol customized according to the patient's time-attenuation response. Heart Vessels 29:149–155
25. Martin SS, Giovagnoli DA, Abadia AF, et al (2020) Evaluation of a Tube Voltage–Tailored Contrast Medium Injection Protocol for Coronary CT Angiography: Results From the Prospective VOLCANIC Study. Am J Roentgenol 215:1049–1056. https://doi.org/10.2214/AJR.20.22777
26. Yin W-H, Yu Y-T, Zhang Y, et al (2020) Contrast medium injection protocols for coronary CT angiography: should contrast medium volumes be tailored to body weight or body surface area? Clin Radiol 75:395-e17

Supplementary material

Analysed classifiers

LR is the standard technique for clinical models. LR is a linear model and is not able to handle the nonlinearities in the data properly. The RF and XGB are an ensemble of decision trees, with different approaches, leading to more complex models but these models can deal with the non-linearity. The SVM and NN are traditional and well-established ML techniques and are also able to deal with the non-linearity of the data differently: the SVM uses hyperplanes in a multidimensional space and the NN uses multiple hidden layers with non-linear activation functions.

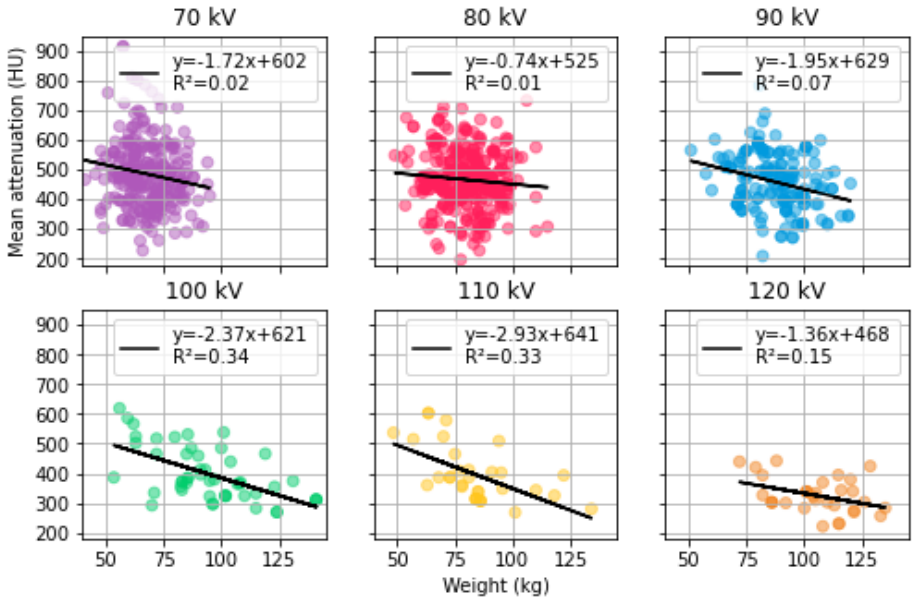

Figure I. Scatter plot showing the relationship between mean attenuation and patient weight per kV group.

Table I. Hyperparameters grid used for SVM.

Classifier	Kernel type	Penalty parameter C	Kernel coefficient γ	Degree of the Polynomial kernel	Class weight
SVM	Linear	[0.1, 1, 10, 100, 1000]	n.a.	n.a.	[None, Balanced]
	Radial basis function	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	n.a.	[None, Balanced]
	Polynomial	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	[2, 3, 4]	[None, Balanced]
	Sigmoid	[0.1, 1, 10, 100, 1000]	[1, 0.1, 0.01, 0.001]	n.a.	[None, Balanced]

Table II. Hyperparameters grid used for RF, GB, and NN.

Classifier	Parameter name	Parameter value
RF	Number of trees	[50, 100, 200, 500]
	Max features	[auto]
	Max depth	[2, 4, 8]
	Min samples per split	[2, 4, 8]
	Min samples per leaf	[1, 2, 4]
	Class weight	[None, Balanced]
XGB	Number of trees	[1000]
	Max depth	[3, 6, 12]
	Gamma	[0, 1, 5, 10]
	Subsample	[0.9, 0.7]
	Learning rate	[0.1, 0.05]
	Col sample by tree	[1, 0.7]
	Min child weight	[1, 5]
	Early stopping rounds	10
	Eval set size	0.1
	Scale pos weight	[1, 2, 3, 4]
NN	Activation	[relu]
	Hidden layer sizes	[50], [100], [10, 10], [50, 50], [5, 5, 5], [10, 10, 10], [50, 50, 50]
	Alpha	[0.001, 0.0001]
	Solver	[adam]
	Learning rate	[adaptive]
	Initial learning rate	[0.1, 0.01, 0.001]
	Early stopping	True
	N iter no change	10
LR	Penalty	[L1, L2]
	Solver	[Liblinear, lbfgs]
	C	[0.1, 1, 10]

Table III. Average LR coefficients over folds.

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Peak height - test bolus (HU)	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Time to peak - test bolus (s)	0.14	0.20	0.17	0.13	0.14	0.15	0.14	0.16	0.17	0.17	0.16
Time to start - test bolus (s)	-0.09	-0.13	-0.10	-0.07	-0.09	-0.11	-0.05	-0.10	-0.10	-0.14	-0.10
Age (yrs)	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02	-0.02	-0.02
Height (cm)	0.06	0.07	0.05	0.06	0.05	0.05	0.05	0.06	0.06	0.06	0.06
Weight (kg)	0.03	0.02	0.03	0.02	0.02	0.03	0.03	0.03	0.03	0.02	0.03
Average heart rate (bpm)	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.01	0.02
Iodine delivery rate (g I/s)	0.55	1.74	0.94	0.83	0.69	0.42	1.15	1.80	0.40	1.58	1.01
Total iodine load (g)	-0.54	-0.77	-0.60	-0.60	-0.60	-0.62	-0.65	-0.80	-0.73	-0.68	-0.66
Sex (female)	0.14	0.05	-0.13	-0.05	-0.18	0.01	0.05	-0.06	-0.08	0.12	-0.01
Tube voltage 70 kV	-2.40	-4.45	-2.40	-3.79	-2.52	-3.86	-2.48	-4.31	-3.19	-3.81	-3.32
Tube voltage 80 kV	-2.00	-3.82	-2.10	-3.45	-2.12	-3.37	-2.06	-3.72	-2.57	-3.53	-2.87
Tube voltage 90 kV	-0.79	-2.21	-0.73	-1.93	-0.72	-1.88	-0.82	-2.08	-0.79	-2.10	-1.41
Tube voltage 100 kV	1.07	0.07	1.12	0.27	1.31	0.22	1.24	0.00	1.45	0.00	0.68
Tube voltage 110 kV	1.06	0.00	1.38	0.00	1.17	0.00	1.47	0.16	1.78	0.00	0.70
Tube voltage 120 kV	2.61	3.04	2.54	1.65	2.89	1.84	2.56	2.53	3.16	1.61	2.44

8

Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning

An application to Phospholamban p.Arg14del mutation carriers

Lopes RR, Bleijendaal H, Ramos LA, Verstraelen TE, Amin AS, Wilde AAM, Pinto YM, de Mol BA, Marquering HA.

Computers in Biology and Medicine. 2021 Apr 1;131:104262.

DOI: 10.1016/j.compbiomed.2021.104262

Abstract

The pathogenic mutation p.Arg14del in the gene encoding Phospholamban (PLN) is known to cause cardiomyopathy and leads to increased risk of sudden cardiac death. Automatic tools might improve the detection of patients with this rare disease. Deep learning is currently the state-of-the-art in signal processing but requires large amounts of data to train the algorithms. In situations with relatively small amounts of data, like PLN, transfer learning may improve accuracy. We propose an ECG-based detection of the PLN mutation using transfer learning from a model originally trained for sex identification.

The sex identification model was trained with 256,278 ECGs and subsequently finetuned for PLN detection (155 ECGs of patients with PLN) with two control groups: a balanced age/sex matched group and a randomly selected imbalanced population. The data was split in 10 folds and 20% of the training data was used for validation and early stopping. The models were evaluated with the area under the receiver operating characteristic curve (AUROC) of the testing data. We used gradient activation for explanation of the prediction models.

The models trained with transfer learning outperformed the models trained from scratch for both the balanced (AUROC 0.87 vs AUROC 0.71) and imbalanced (AUROC 0.0.90 vs AUROC 0.65) population. The proposed approach was able to improve the accuracy of a rare disease detection model by transfer learning information from a non-manual annotated and abundant label with only limited data available.

Introduction

The phospholamban (PLN) p.Arg14del mutation is known to cause both arrhythmogenic – and dilated cardiomyopathy in patients with this condition leading to increased risk of sudden cardiac death and end-stage heart failure (1). This frequently necessitates implantation of an implantable cardioverter defibrillator (ICD) or even heart transplantation. PLN p.Arg14del is a rare mutation, present in 0.08%- 0.38% in selected cardiomyopathy cohorts (2) and the early diagnosis of this mutation, especially before patients get symptomatic, could potentially prevent sudden cardiac death, for example by implanting an ICD (3,4). The current standard diagnosis is performed by highly specialized electrophysiologists and cardio-geneticist with DNA sequencing, the latter being time-consuming and expensive. Alternatively, PLN mutation diagnostics could benefit from tools for automatic identification of patients with the mutation on a large scale and low cost, for example on ECGs. Some characteristics in electrocardiograms (ECG), like inverted T waves in leads V_4 to V_6 and low-voltage ECGs, can often be identified in patients with this PLN mutation (3,5–8).

Electrocardiography is a commonly used, non-invasive, and low cost inexpensive tool to assess electrical activity of the heart and is used to identify heart rhythm irregularities and other related cardiovascular diseases (9). Besides disease related information, an ECG carries a lot of patient-specific information since it is a unique and distinctive combination of signals, which differ per individual (10). It has been shown that ECG can also be used to identify sex (11) or even the effects of intake illegal drugs (12). Recently, artificial intelligence has emerged as a powerful tool and has been broadly applied in cardiology, leading to promising results in multiple diagnostics related tasks, including the analyses of ECGs (8,13–18)**.**

Advances in artificial intelligence, specifically the developments of deep learning (DL) architectures and convolutional neural networks (CNNs), mostly rely on convolutional filters to extract features from data. Based on these advances, multiple studies have been performed to automatically recognize various diseases, like arrhythmia, coronary artery disease and genetic mutations in ECG signals (8,13–15). A recent study introduced a CNN-based method that was able to estimate the age and sex of patients based on the entire ECG (16). Different

from traditional Machine Learning techniques, which rely on handcrafted features, CNNs have the capability of extracting temporal and spatial (variation of the signal over the different leads) information from raw data, as illustrated in Figure 1, without explicit definitions of the features searched for. In general, CNNs need large amounts of (manually annotated) data to train and to achieve accurate results. These large amounts of annotated data are commonly not available, especially for tasks in rare diseases. To deal with a limited amount of training data, it has been proposed to pre-train models on a different domain or task and use these models as the initial step to subsequently adapt them for another task (19,20). This approach, known as transfer learning, aims to take advantage of the parameters learned from an initial task, with more labelled data available, commonly providing a better starting point for further training on a different task. During the training phase of DL models, kernels are learned automatically and are used to extract informative features of the data. This process of learning the kernels might not be optimal if, for example, the amount of data is low, leading to redundant or non-useful kernels. Using transfer learning, the kernels are learnt during an easier task or from a problem with more data, leading to better results when finetuning it to specific tasks. Transfer learning has previously been applied for ECG for arrhythmia detection, transferring information from models pre-trained on a large database of natural images (21) and for pre-training on human ECGs to improve equine ECG classification tasks (22).

A previous study (8), demonstrated that it is possible to detect PLN in ECG signals using machine learning techniques and a beat-to-beat approach. The models outperformed experts in terms of sensitivity and accuracy and had a reasonable evaluation. However, it should be noted that this accuracy is not yet high enough to be used in clinical practice. Moreover, it should be considered that this score was achieved in a balanced dataset, which contained the same number of PLN and non-PLN patients, which is not representative for real-life scenarios given the rarity of the mutation.

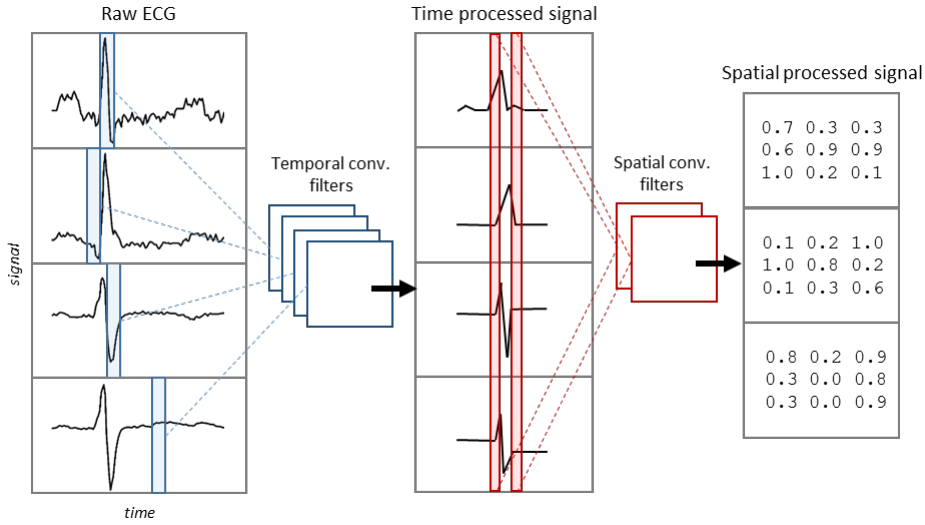

Figure 1. Illustration of layers with convolutional filters to extract temporal (blue) and spatial (red) information from ECGs with multiple leads. A CNN can be composed of many of these layers. After many layers, the signal becomes less human interpretable.

PLN-mutation is also hampered with a small amount of data because of the sparsity of this disease. To potentially improve the accuracy of an ECG-based detection of PLN gene mutation using DL, we propose using transfer learning to deal with the rarity of this disease. We aimed to take advantage of the large amount of non-labelled data by creating a pre-trained model. As the available ECGs have been acquired for multiple reasons, we created the pre-trained model of demographics information since this kind of information was available for all patients. Predicting demographics has been performed before [16], which motivated us to follow this approach. In our study, we have selected "sex – classification" instead of "age – regression" because the of the dichotomous classification of PLN identification. Note that such a pre-trained model could be used for multiple medical classification tasks. We evaluate the potential benefits of using transfer learning, by pre-training a CNN model on a large ECG database, using the entire signal, analyzing whether subsequently finetuning this model results in increased accuracy compared to training it from scratch. To this end, we 1) pre-train a CNN model for sex identification, for which no manual or expensive annotations are needed, as the source task and fine-tune this model for the identification of PLN patients, 2) evaluate the model using balanced and

unbalanced datasets, and 3) identified the parts of the ECG's valuable for the prediction to allow the possibly identification of novel patterns in the ECGs.

Methods

We introduce a two-step approach in which we first build a CNN model for sex identification using ECGs, which is subsequently used as pre-trained model to further develop the model to identify PLN patients. As matters of comparison, we trained the models from scratch using the same architecture. A gradient-based technique is used for the visualization and interpretations of the predictions.

Sex database

Single ECGs from a total of 256,278 patients were included (52% male and mean age of 50 years). We included patients, from 18 to 60 years, with at least one digital 8-leads ECG available with 10 s of duration and a 500 Hz or 250 Hz sampling rate. For patients with multiple ECGs available, only the first ECG acquired was included. The ECGs were acquired with a GE ECG machine and stored using the MUSEweb data management system (GE Healthcare, Chicago, Illinois, United States of America). The ECGs were resampled to 250 Hz, leading all ECGs to have the same length, with 8 leads x 2500 sampling, for further feeding the CNN. Only the 8 main leads (I, II, V1-V6) were included since the others (III, aVR and aVL) are derived leads and including those might add redundancy to the model.

Development of the sex identification model

The architecture and hyperparameters for the sex identification model were the same as the ones proposed by (16). The network, summarized in Figure 2, is composed of blocks with convolutional kernels, batch normalization, and max pooling. The convolutional kernels of the first 8 blocks are of size (1, X), with X decreasing from 7 to 3, and filters over each individual lead. In the 9^{th} block, the convolutional kernels are of size (8, 1) and filters over all the leads together. The two final blocks are composed of fully connected layers.

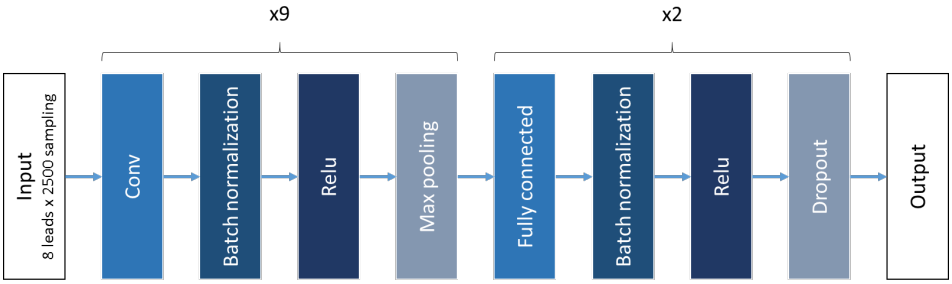

Figure 2. Schematic representation of the architecture for the sex identification and PLN detection. The first 8 blocks process time signal and the subsequently one is filtering over the leads. The model was initially trained on a larger database for sex identification and further finetuned for PLN detection.

Early stop of the training was triggered during training if the validation error of the network increased for five epochs. Three different learning rates (1e-2, 1e-3 and 1e-4) and four different batch sizes (16, 32, 64, 128) were tested and the one showing the lowest validation loss was selected. The model was trained with Adam optimizer, a learning rate of 1e-3 and batch size of 32. We used the default values for the other hyperparameters. An initial split of the data was performed for training (80% of patients) and test (20% of patients). Besides that, 20% of the training data was used as the validation set.

PLN database

The PLN dataset and the data collection protocol are the ones presented by Bleijendaal et al. (8). The ECGs were classified as PLN or control. Similar to the sex database, single ECGs with 10 s were included. Differently from the approach presented in Ref. (8), the entire signal was used and all the ECGs were resampled to 250 Hz to have the same length as the sex database (8 × 2500). Figure 3 displays an example of ECGs of a PLN patients and a control patient as well as some of the known PLN features in the ECG.

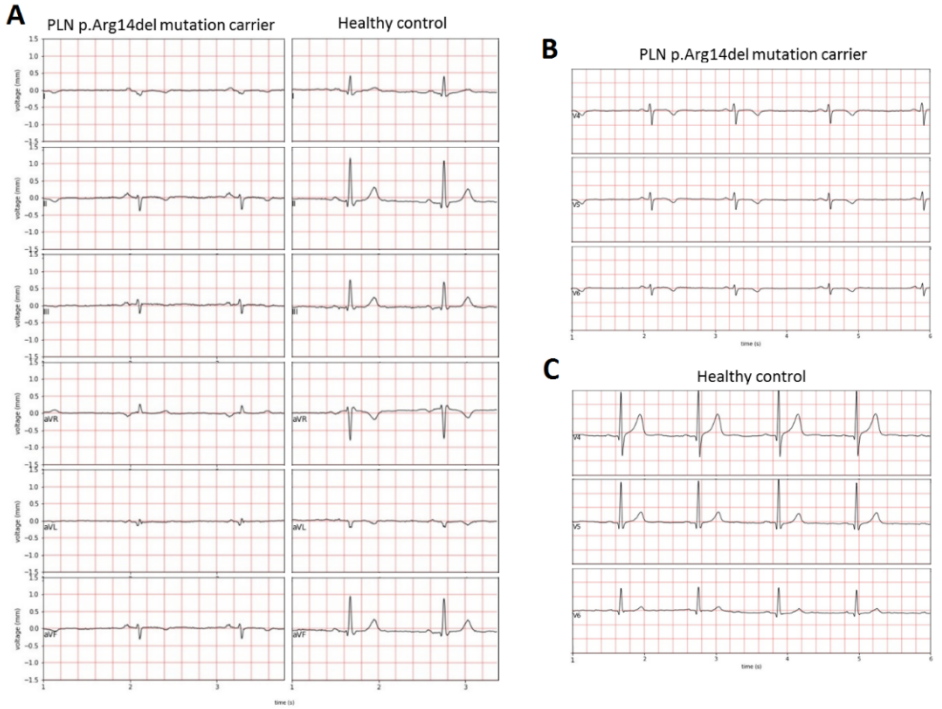

Figure 3. Example of a PLN p.Arg14del mutation and healthy control ECG patients used on this study. ECG features associated with this mutation such as low QRS voltages on the extremity leads of the ECG (A) and T-wave inversion on the lateral leads V4-V6 (B and C) are shown.

Two different experiments were performed with *balanced* and *imbalanced* populations. The imbalanced control dataset was developed using patients, aged from 18 to 60 years old, who underwent general clinic ECG acquisition (non-cardiovascular pre-operative screening at the out-patient clinic) of the Amsterdam UMC, location AMC. From this dataset, with 13,467 patients, a matched control group was selected to create the balanced experiment, so both control and PLN group have similar distribution regarding age and sex. For the imbalanced experiment, all 13,467 patients were included without any constraints regarding the selection. All baseline characteristics of the patients used in our models can be found in Table 1. To avoid a possible bias, all PLN and control patients were not included in the sex identification experiment.

Development of the PLN model

In both experiments, the dataset was split into training, validation and testing set. The training and validation sets were used to train and optimize the models, and the test set was kept unseen until the final evaluation. Because of the small number of patients with the PLN condition and to have a more robust evaluation, a stratified shuffle split with 10 folds was applied as cross-validation for both experiments.

Table 1. Characteristics of the PLN patients, balanced control group and imbalanced control group. Values are represented as median and interquartile ranges, unless stated otherwise. *bpm* = beats per minute, *ms* = millisecond

Variable name	PLN	Control balanced	p-value	Control imbalanced	p-value
n	155	155	1.000	13,467	
Age	39 [28–50]	39 [28–50]	1.000	52 [45-57]	<0.001
Sex (male %)	63 (41)	63 (41)	1.000	6207 (46)	0.225
Ventricular rate (bpm)	68 [60–75]	65 [57–73]	1.000	68 [60-77]	0.071
Atrial rate (bpm)	68 [60–75]	66 [57–73]	1.000	68 [60-78]	0.193
QRS duration (ms)	86 [80–94]	94 [84–104]	1.000	90 [82-98]	0.037
QT interval (ms)	388 [368–406]	400 [374–426]	1.000	390 [370–414]	0.535
QT corrected (ms)	407 [394–424]	410 [401–429]	1.000	412 [403–427]	0.113

The transfer learning PLN model was built using the pre-trained model for sex identification. The last layer of the pre-trained model, responsible for the classification of the sex, was discarded and a new one was added. All layers of the model remained trainable. Similar to the sex identification model, a validation set (20% of the training data) was used for the analysis of the network convergence. Also, during the finetuning of the model, early stopping in case of increasing validation error was performed and multiple learning rates and batch sizes were tested. The model was optimized using RMSprop (23), similarly to (8). To deal with the high disproportion of class samples in the imbalanced experiment, class weights were assigned during the training, so the class with fewer samples have a higher impact during the training process. With this approach, the mistakes committed by the model on PLN samples are highly

penalized when compared to mistakes on the control group, compensating the class imbalance.

Model evaluation

We evaluated the models using the average Area Under the Receiver Operating Characteristic curve (AUROC) with its 95% confidence interval (CI). Besides that, the specificity, sensitivity and area under Precision-Recall curves (AUPRC) were reported. The AUPRC is more informative than ROC curve and (24) commonly used for detection of rare diseases (25). Unless stated otherwise, all reported results are based on the test set, which was kept unseen by the model until the evaluation. To check whether the difference in AUROC between the models was statistically significant, the Wilcoxon signed-rank test was performed for each experiment.

Prediction model interpretation

For a better understanding of the prediction models, we applied Grad-CAM (26). This technique uses the gradient of the model's classification to localize important regions of the ECG. We applied Grad-CAM for all patients from one of the balanced test sets (n = 62). Only one test set was selected to avoid repetition on the evaluated ECGs. Important regions determined with Grad-CAM were qualitatively assessed by a single investigator (H.B.) and classified as either containing actual relevant information or the presence of ECG noise (e.g. baseline shift or other noise) that might have induced a high gradient response in the model. To analyse which leads contain the most important information, we calculated the sum of importance per lead for all assessed ECGs.

Results

The sex identification model achieved an accuracy of 0.82 in the validation and 0.83 in the test set. For the PLN experiment, the transfer learning model achieved an average AUROC of 0.87 (95% CI: 0.84-0.90) in the balanced experiment, compared to an AUROC of 0.72 (95% CI: 0.66-0.78) for the model trained from scratch, with a statistically significant difference (p-value<0.01). Regarding the imbalanced experiment, again the transfer learning model (AUROC: 0.90, 95% CI: 0.87-0.93) outperformed the model created from scratch (AUROC: 0.65, 95%

CI: 0.56-0.74) with a statistically significant difference (p-value<0.01). In Figure 4 we show the ROC of the models with the balanced and imbalanced datasets.

Table 2 shows the AUROC, AUPRC, sensitivity and specificity of all developed models. For comparison, we also present the results of a previous study, where AUROC of 0.80 was achieved by a machine learning model and 0.65 by a specialist.

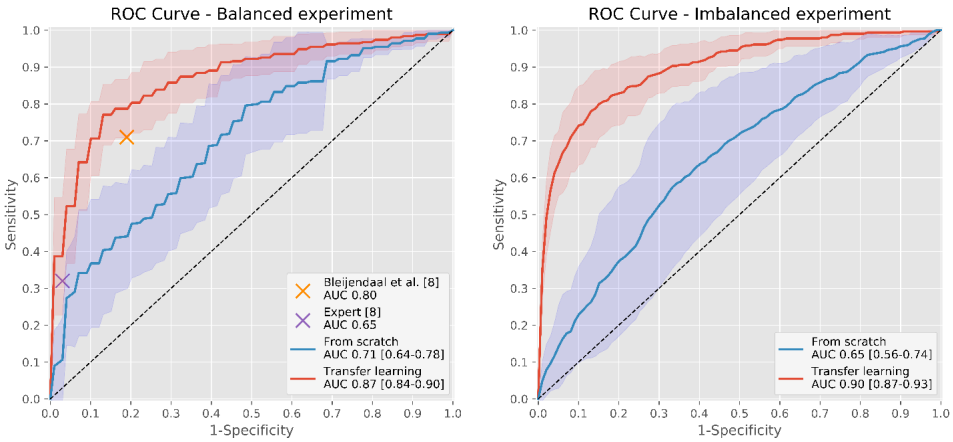

Figure 4. Average ROC curve and 95% CI for PLN detection in a balanced dataset (left) and imbalanced dataset (right) with transfer learning (red) and the model trained from scratch (blue). Results from a previous study (8) are represented with an "X". *ROC* = receiver operating characteristic, *AUC* = area under the curve

Figure 5 shows the Precision-Recall Curve for balanced and imbalanced experiments. In both experiments, the transfer learning models (AUPRC of 0.88 and 0.28) outperformed the models trained from scratch (AUPRC of 0.70 and 0.03).

In Figure 6, we present two examples of ECGs and regions of maximum gradient activation (most relevant for prediction). In both ECGs the QRS complex is the region with a large gradient activation, also activation in the region of the T-wave can be seen. From the total 62 ECGs that were analyzed, the ECG region with most activations, for both PLN and control, was the QRS complex (27 (87%) for both PLN and control). In 2 ECGs (6%) of the control group, the T-wave was the region with most activation. According to the assessed ECGs, the most important

lead is the V2 for both PLN and control ECGs. A summary of the ECG regions that were found to be the most relevant for PLN detection is presented in Table 3.

Table 2. Evaluation of the models trained from scratch and with transfer learning in balanced and imbalanced dataset. Both sensitivity and specificity were measured with 0.5 as threshold. Not available information is represented as *na*.

Model/Metric	Sensitivity	Specificity	AUROC	AUPRC
Balanced				
From scratch	0.36	0.81	0.71	0.70
Transfer learning	0.80	0.78	0.87	0.88
Bleijendaal et al. (8)	*0.71*	*0.81*	*0.80*	*na*
Expert (8)	*0.32*	*0.97*	*0.65*	*na*
Imbalanced				
From scratch	0.81	0.45	0.65	0.03
Transfer learning	0.63	0.95	0.90	0.28

AUROC Area Under the Receiver Operating Characteristic curve, *AUPRC* Area Under the Precision-Recall curves.

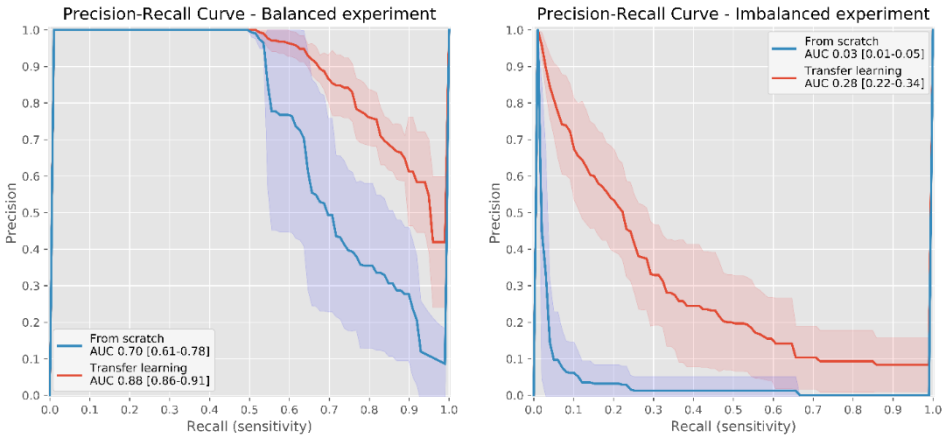

Figure 5. Precision-Recall curves and 95% CI for balanced (left) and imbalanced (right) datasets with transfer learning (red) and the model trained from scratch (blue). *AUC* = area under the curve

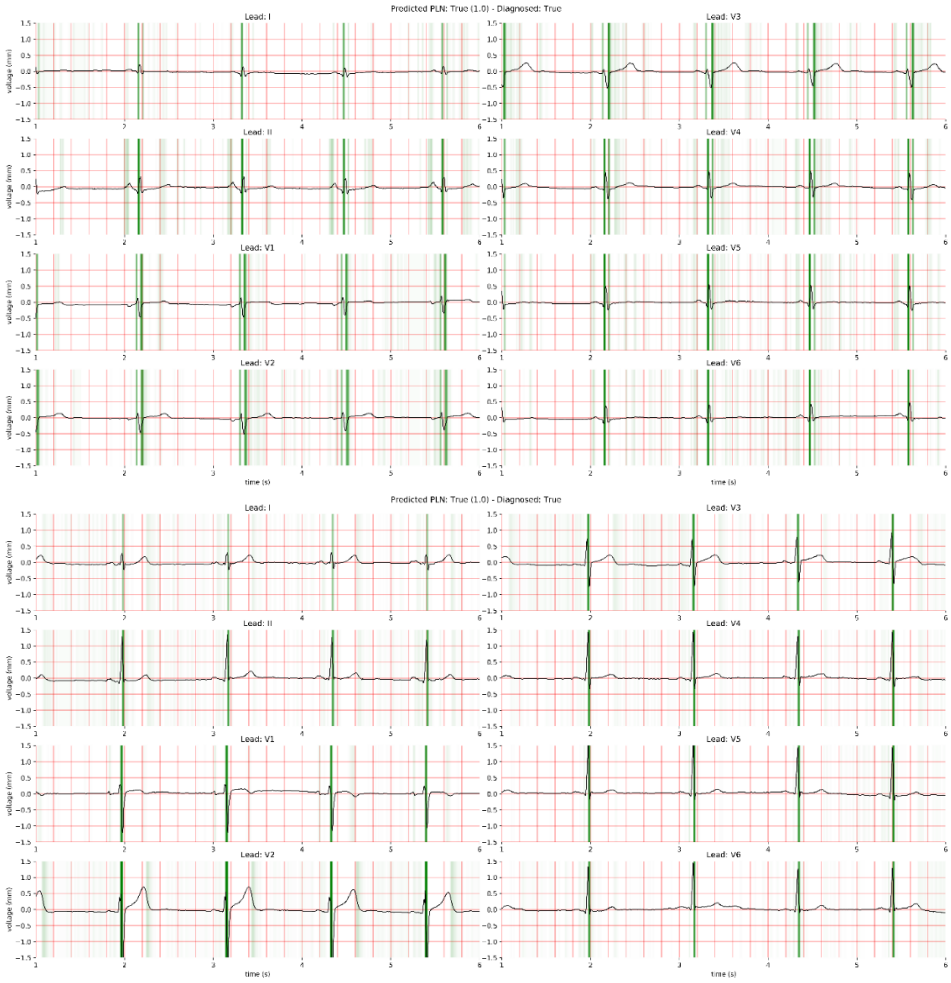

Figure 6. Examples of positive cases in which Grad-CAM highlighted the QRS complex, specifically the upward leg of the R and S-waves (top) the downward leg of both the R and S-waves (bottom). Regions highlight in green are the most relevant for the model.

Discussion

In this study, we have shown that pre-trained DL models can significantly improve ECG-based PLN patient detection. The improvement was observed in AUROC and AUPRC measures for both the balanced and imbalanced experiment. Our results confirm that transfer learning, even from simple tasks like sex identification, could be beneficial to models trained to predict diseases with considerable small datasets.

Table 3. This table shows the regions of the electrocardiogram and leads with most gradient activation (using Grad-CAM), for both PLN and the control in a subset of our data.

	PLN (n = 31)	Control (n = 31)
Correct prediction by CNN (%TP/%TN)	22 (71)	15 (48)
QRS complex (%)	27 (87)	27 (87)
Q-wave (%)	*0 (0)*	*0 (0)*
R-wave (%)	*3 (10)*	*0 (0)*
S-wave (%)	*7 (22)*	*7 (22)*
Downward leg of R and S (%)	*8 (26)*	*14 (45)*
Upward leg of R and S (%)	*9 (29)*	*6 (19)*
T-wave (%)	0 (0)	2 (6)
Other than QRS-complex or T-wave (%)	1 (3)	1 (3)
Artifact/noise (%)	3 (10)	1 (3)
Leads (%)*		
I (%)	2 (6)	2 (6)
II (%)	5 (16)	0 (0)
V1 (%)	1 (3)	4 (13)
V2 (%)	11 (35)	8 (26)
V3 (%)	4 (13)	7 (22)
V4 (%)	3 (10)	5 (16)
V5 (%)	5 (16)	2 (6)
V6 (%)	0 (0)	2 (6)

CNN Convolutional Neural Network*, TP* True positive*, TN* True negative, *one sample did not return gradient.

A recent study by Bleijendaal et al. (8), presented the first approach to detect PLN from ECGs using DL from individual heart beats on ECG and showed reasonable accuracies when compared to specialists. The presented transfer learning model in this study, using the entire signal, outperformed both their DL models and the results achieved by experts. It should be noted that in the study from Bleijendaal et al. [8], the training and testing were only conducted using a balanced dataset, which is not representative of clinical practice given the rarity of the disease. Besides that, while we decided to use a previous evaluated architecture for sex identification as proposed by Attia et al. (16), they used a beat-wise approach and a reasonably shallow architecture developed specifically their study given the

limited amount of data. This shallower architecture might be the reason why their model outperformed our model trained from scratch.

Many other studies that applied transfer learning for cardiac disease have shown promising results. Kachuee et al. (27), for instance, presented a CNN model trained for arrhythmia classification and finetuned for myocardial infarction classification. Other studies focused on transfer learning from "off-the-shelf" pre-trained models on natural images: Salem et al. (21) outperformed traditional ML models by finetuning a model with spectrograms from ECGs and Xiao et al. (28) took advantage from a pre-trained model for early detection of myocardial ischemia on ambulatory ECGs. In the current active field of deep learning, many studies aiming for optimizing architectures have been published. Regular CNNs as well as hybrid approaches, using CNN and Recurrent neural networks (RNN), are promising approaches and adequate to time series as presented by Hong et al. Recurrent neural networks have a different kind of architecture capable of learning long-term temporal dependencies. Petmezas et al. (29) developed a hybrid CNN-RNN pipeline to detect atrial fibrillation. Londhe and Atulkar (30) presented a deep learning architecture for segmentation of the P, QRS and T waves. However, such networks are more complex and less efficient compared to regular CNNs. On the other hand, Yan et al. (31) presented an approach using spiking neural networks, which has a lower energy cost and showed similar results to the traditional CNN. Since the aim of our study was to evaluate the added value of transfer learning, the comparison with other DL approaches was beyond the scope of our study.

With the higher accuracy of the presented approach compared to previous efforts, the potential value of this application in clinical practice has improved. Moreover, the assessment of the accuracy in a population that better resembles clinical populations compared to previous efforts, allows a better estimation of its value in clinical practice. The presented approach includes a trade-off between sensitivity and specificity, which can and should be adjusted for the desired application. With higher precision (positive predictive value), this model could be deployed to identify possible PLN candidates in the clinic. However, before such an application can be integrated in practical environments, this approach has to be more extensively tested in multiple populations. Next, various aspects such

as the clinical value, needs to be assessed. Besides that, the same approach could be used to improve detection of other diseases, such as Long-QT syndrome, or even detect multiple diseases with a similar model using a multi-class prediction model, as it has been proposed by Ribeiro et al. (32). However, since other mutations are less malignant and/or less common in the Netherlands, we did not have large quantities of data for other mutations. Although the model trained with imbalanced data showed high AUROC values, it can be overoptimistic for the imbalanced experiment given the high class imbalance of the experiment. While the sensitivity only takes in account the positive samples, the specificity takes in account the negative samples and even a small percentage of false positives leads to a number reasonable higher than the positive samples. Evaluation with other metrics, such as the PR-Curves, illustrates the effect of the imbalance in the results.

This study suffered from some limitations. A small quantity of the ECGs had noisy regions and yielded high activations on the prediction interpretation. Nevertheless, the number of ECGs with most activation on the noisy region was relatively low (6% of the assessed samples), which is very unlikely to significantly affect the models. Besides that, the PLN dataset is rather small, comes from a single center only and it is not guaranteed that subjects in the control group, for both balanced and imbalanced experiments, do not have PLN (since they were not tested). We assume that the control subjects do not have PLN because of the rare nature of the disease. In addition, the imbalanced control population is reasonably older than the PLN population and, although we visualized PLN related regions in the ECGs using the gradient technique, the model might be using some age-related features. Validation in an external dataset (preferably from a different center) is still necessary to evaluate how generalizable our models are to a different population. Moreover, we suffered from the “black box” effect and even with an advanced approach for prediction interpretation, it is not possible to completely understand and interpret on what ECG features the predictions are performed for the PLN identification. While our model is mainly influenced by the QRS complex, Bleijendaal et al. (8) found the T-wave as an important region for their model using a different method. Our model is mainly activated by either the upward leg of the R and S waves or the downward leg of the R and S waves. This part of the QRS complex might be

relative to the amplitude of the wave and it is a known PLN feature. This difference is not surprising since both regions are known to be influenced by PLN. Therefore, differences in the methods adopted (transfer learning approach, kernel size and convolution over leads) can result in different learned convolutional kernels to solve a similar classification problem. In this study, we used a previous evaluated architecture for sex identification, but deeper and more complex architectures could be considered, such as a residual network or RNN. Deeper networks, trained with large amounts of data, might lead to robust kernels which can improve prediction capability.

Conclusion

In our study, the accuracy of the ECG-based identification of the rare phenomenon of PLN strongly improved by using a transfer learning approach in which a model was pre-trained on sex classification. Improvement was observed both in balanced and imbalanced experiments. The QRS complex was found to be the most important region in the ECG for PLN identification. We conclude that we can improve the accuracy of the detection of a rare disease by creating a model exploiting transfer learning and using information without the need of manual annotation when only limited data as available.

References

1. Haghighi K, Kolokathis F, Gramolini AO, Waggoner JR, Pater L, Lynch RA, et al. A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. Proc Natl Acad Sci. 2006;103(5):1388–93.
2. Hof IE, van der Heijden JF, Kranias EG, Sanoudou D, de Boer RA, van Tintelen JP, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. Netherlands Hear J. 2019;27(2):64–9.
3. van Rijsingen IAW, van der Zwaag PA, Groeneweg JA, Nannenberg EA, Jongbloed JDH, Zwinderman AH, et al. Outcome in Phospholamban R14del Carriers: Results of a Large Multicentre Cohort Study. Circ Cardiovasc Genet [Internet]. 2014 Aug 1;7(4):455–65. Available from: http://circgenetics.ahajournals.org/cgi/doi/10.1161/CIRCGENETICS.113.000374
4. Towbin JA, McKenna WJ, Abrams DJ, Ackerman MJ, Calkins H, Darrieux FCC, et al. 2019 HRS expert consensus statement on evaluation, risk stratification, and management of arrhythmogenic cardiomyopathy. Hear Rhythm [Internet]. 2019 Nov;16(11):e301–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1547527119304382
5. Posch MG, Perrot A, Geier C, Boldt L-H, Schmidt G, Lehmkuhl HB, et al. Genetic deletion of arginine 14 in phospholamban causes dilated cardiomyopathy with attenuated electrocardiographic R amplitudes. Hear Rhythm [Internet]. 2009 Apr;6(4):480–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1547527109000484
6. van der Zwaag PA, van Rijsingen IAW, Asimaki A, Jongbloed JDH, van Veldhuisen DJ, Wiesfeld ACP, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: evidence supporting the concept of arrhythmogenic cardiomyopathy. Eur J Heart Fail [Internet]. 2012 Nov;14(11):1199–207. Available from: http://doi.wiley.com/10.1093/eurjhf/hfs119
7. Cheung CC, Healey JS, Hamilton R, Spears D, Gollob MH, Mellor G, et al. Phospholamban cardiomyopathy: a Canadian perspective on a unique population. Netherlands Hear J [Internet]. 2019 Apr 26;27(4):208–13. Available from: http://link.springer.com/10.1007/s12471-019-1247-0
8. Bleijendaal H, Ramos LA, Lopes RR, Verstraelen TE, Baalman SWE, Oudkerk Pool MD, et al. Computer versus Cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing phospholamban (PLN) p.Arg14del mutation on ECG? Hear Rhythm [Internet]. 2020 Sep; Available from: https://linkinghub.elsevier.com/retrieve/pii/S1547527120308614
9. De Bacquer D, De Backer G, Kornitzer M, Blackburn H. Prognostic value of ECG findings for total, cardiovascular disease, and coronary heart disease death in men and women. Heart. 1998;80(6):570–7.
10. Pinto JR, Cardoso JS, Lourenço A, Carreiras C. Towards a continuous biometric system based on ECG signals acquired on the steering wheel. Sensors. 2017;17(10):2228.
11. Malik M, Hnatkova K, Kowalski D, Keirns JJ, van Gelderen EM. QT/RR curvatures in healthy subjects: sex differences and covariates. Am J Physiol Circ Physiol. 2013;305(12):H1798–806.

12. Hossain SM, Ali AA, Rahman MM, Epstein EED, Kennedy A, Preston K, et al. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In: IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks. IEEE; 2014. p. 71–82.
13. Isin A, Ozdalili S. Cardiac arrhythmia detection using deep learning. Procedia Comput Sci. 2017;120:268–75.
14. Acharya UR, Fujita H, Lih OS, Adam M, Tan JH, Chua CK. Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. Knowledge-Based Syst. 2017;132:62–71.
15. Aston PJ, Lyle J V, Bonet-Luz E, Huang CLH, Zhang Y, Jeevaratnam K, et al. Deep Learning Applied to Attractor Images Derived from ECG Signals for Detection of Genetic Mutation. In: 2019 Computing in Cardiology (CinC). IEEE; 2019. p. 1–4.
16. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circ Arrhythmia Electrophysiol. 2019;12(9):e007284.
17. Mincholé A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nat Med. 2019;25(1):22–3.
18. Lyon A, Mincholé A, Martínez JP, Laguna P, Rodriguez B. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. J R Soc Interface. 2018;15(138):20170821.
19. Chen L, Xu G, Zhang S, Kuang J, Hao L. Transfer Learning for Electrocardiogram Classification Under Small Dataset. In: Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. Springer; 2019. p. 45–54.
20. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proc IEEE. 2020;
21. Salem M, Taheri S, Yuan J. ECG arrhythmia classification using transfer learning from 2-dimensional deep CNN features. In: 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE; 2018. p. 1–4.
22. Van Steenkiste G, van Loon G, Crevecoeur G. Transfer Learning in ECG Classification from Human to Horse Using a Novel Parallel Neural Network Architecture. Sci Rep. 2020;10(1):1–12.
23. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on. 2012;14(8).
24. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Brock G, editor. PLoS One [Internet]. 2015 Mar 4;10(3):e0118432. Available from: https://dx.plos.org/10.1371/journal.pone.0118432
25. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol [Internet]. 2015 Aug;68(8):855–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435615001067
26. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conf Appl Comput Vis [Internet]. 2017 Oct 30;839–47. Available from: https://ieeexplore.ieee.org/document/8354201/

27. Kachuee M, Fazeli S, Sarrafzadeh M. Ecg heartbeat classification: A deep transferable representation. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2018. p. 443–4.
28. Xiao R, Xu Y, Pelter MM, Mortara DW, Hu X. A deep learning approach to examine ischemic ST changes in ambulatory ECG recordings. AMIA Summits Transl Sci Proc. 2018;2018:256.
29. Petmezas G, Haris K, Stefanopoulos L, Kilintzis V, Tzavelis A, Rogers JA, et al. Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets. Biomed Signal Process Control [Internet]. 2021 Jan;63:102194. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1746809420303323
30. Londhe AN, Atulkar M. Semantic segmentation of ECG waves using hybrid channel-mix convolutional and bidirectional LSTM. Biomed Signal Process Control [Internet]. 2021 Jan;63:102162. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1746809420303049
31. Yan Z, Zhou J, Wong W-F. Energy efficient ECG classification with spiking neural network. Biomed Signal Process Control [Internet]. 2021 Jan;63:102170. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1746809420303098
32. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun [Internet]. 2020 Dec 9;11(1):1760. Available from: http://www.nature.com/articles/s41467-020-15432-4

9

Discussion

General Discussion

The presented research is characterized by optimization, testing and validation of machine learning (ML) models and the discussion of their clinical usefulness compared to standard approaches and technologies. This research required close collaboration with physicians to understand the context and problems, get access to datasets, and define boundary conditions regarding model development, validation and interpretation. The studies of this thesis have cardiology as a common denominator, but they cover only a small fraction of possible topics in the field. The selected studies, here presented as chapters, were conducted considering the availability of data and clinical needs defined by a clinical contributor. The major objective of this thesis was to develop and validate ML models in the cardiology domain. To this end, I teamed up with several physicians and scientists to understand their problems and how ML could be used to potentially solve them, while aligned with their expectations.

The main findings of this thesis also have the use of ML to enhance cardiology as a common denominator. It is important to note that the robustness and accuracy of the presented ML models depend not only on the technical aspect of the models but mainly on the quality of the data used for their creation. The physicians, aligned with statisticians, commonly want to compare ML with traditional statistical approaches based on linear or logistic regression. The results we obtained show that there is almost no difference in accuracy when models are created with different techniques. Despite that, there are huge accuracy gaps when different feature sets are used.

It is commonly assumed that ML models require a lot of data. The large data demand can be seen either as a drawback or as a strength: the models can learn from a large amount of data and benefit from non-linear relations between the features. The traditional statistical approaches are usually trained based on demographics, tabular clinical data, and scores given by physicians. Unless important non-linear relationships are present in the feature set, it is unlikely that

ML models will significantly outperform the traditional approaches. However, deep learning is seen as state-of-the-art when considering the analysis of multi-dimensional data, such as medical images (1), and its performance has been commonly claimed to outperform clinicians (2). In addition, non-parametric models and micro-simulations (simulating small variations in the data to observe changes in the outcome), can improve decision-making and risk assessment (3).

The availability and low quality of data, including disparities in datasets, are the major limitations of our studies. As an example, it is common to consider information from the census to track if a patient from a cohort has died. In this sense, it is unclear if the patient has died as a consequence of the surgery or was hit by a car. In bi-institutional datasets, one centre's dataset may contain three times as many patients as the second centre. These are a few of many examples that might end up adding noise to the data and jeopardizing its quality. Despite various registries harbouring an abundance of data, the collection and storage of data are not standardized. Using data before the harmonization required for a registry demands additional steps to harmonize data from different centres, which usually means fewer parameters available for modelling of any kind. The standardization issues could be reduced if interoperability standards were adopted and applied regularly in a particular discipline like cardiology. Data harmonization and availability will be addressed in more detail as follows.

The use of ML has been investigated in several areas of cardiology (4–7). The use of ML techniques increased largely in cardiology and various other medical fields due to the advances in storage, computer processing, and a new focus on data quality (8–10). Although not addressed in our studies, today's requirements of data security, data access, and privacy are as challenging as organizing and conducting research with patient datasets itself. A healthcare data storage system, with the guarantee of data security, is now considered a critical requirement. Not only for privacy and security reasons but also because this data is extremely valuable to gain insights, create prediction models, reduce healthcare costs, and mainly, to improve people's lives (11,12). These efforts pay off only when one can be sufficiently certain that the models are accurate and robust enough to be used and that they aggregate value. How they are developed and validated is of utmost importance to guarantee the accuracy, robustness, and stability of the

models. As a starting point, considering that much work had to be done in this respect, we explored ways to assess and deal with these three requirements simultaneously in several applications for cardiology.

Similarly to the experience of other authors, we had to deal with multiple other challenges, such as data sharing limitations, limited amount of available data, and model explanation. These challenges can affect the quality of the model and, therefore, its performance. Although the quality of data must be high enough to ensure reasonable models, it is an illusion that one can work with perfect datasets in terms of correct registration and completeness. Consequently, proper metrics and evaluation must be considered to assess how the models deal with imperfection and to which extent imperfect data affect model performance.

In this thesis, we focused on strategies to improve accuracy by creating models based on different training approaches, such as finetuning, local, and distributed learning. In the studies, we also assessed the robustness and stability of models and used validation strategies such as internal, external, and temporal validation. We found that our applications and outcomes carried acceptable degrees of accuracy and robustness. However, remains room for improvement by exploring other technical approaches. Nevertheless, the quality of the data remains a limiting factor that should be reduced by data-centric approaches to assess and potentially improve it (10,13).

Challenges

Data harmonization

It is a common scenario for data to be acquired over multiple years and collected from different sources. Eventually, some information that is crucial for one medical centre may not even be considered in another. It might occur that, when external validation is being performed, the external data is not on the same standard and format as the data that was used to train the models. For the TAVI mortality prediction models described in Chapters 2 and 3, for instance, some features had to be harmonized with the external dataset used for training and validating the models: some of them were excluded (such as scores from questionnaires on quality of life) for only being available at one of the centres; while others, such as time measurements of the patient's heartbeat (measured in

milliseconds and centiseconds) had to be converted to the same unit of measurement. Also, features available in only one of the centres or with a high percentage of missing values were excluded. As in Chapter 5, the use of registries is one resource to minimize data availability and standardization issues.

Another possible problem could be that an equipment that has been used for multiple years during the generation of the dataset is changed, either in total or just in some settings. Consequently, data used in Chapters 2-4 that was acquired over several years could have been through possible changes that were not assessed. In Chapter 5, we investigated the accuracy and stability of TAVI mortality prediction models over the years using a large national registry. In Chapter 7, we created models to predict insufficient contrast enhancement on Coronary CT Angiography. However, the bolus' volume and injection speed of the test bolus were modified during the data acquisition period. These changes were not taken into consideration and should be assessed in future studies.

Data availability

It is a rule of thumb that the more data you have, the better the ML models are. However, sometimes, it is impossible to have a large quantity of clinical data regarding a specific procedure or disease. This lack of data mainly occurs with rare diseases or complex procedures, such as artificial valve implantation. To deal with the limited amount of data, we used finetuning (Chapters 3 and 8) and distributed learning (Chapter 4) in this thesis. Both approaches are relevant not only because they use more data to train the models but also because they do not require data to be shared among centres, guaranteeing confidentiality for patients. Also shown in Chapter 4, using a stacking method (for instance, training a simple logistic model with the probability output from other models) might be beneficial and achieve higher accuracy than individual models. This approach is easy to be implemented, as only the models are shared between centres and they can be optimized individually in each centre.

Model optimization and validation

When developing ML models, it is a good practice to have a test set that consists of data unseen by the model, to evaluate the trained models. This is highly recommended as many of the complex techniques, such as random forest and

XGBoost, are prone to overfitting. It is usually desired to have all the available data used for testing, validated either by k-fold cross-validation or repeated shuffle split. Also, another good practice is not using validation data, which is the data used to optimize the models' hyperparameters or interrupt the training, to evaluate the models. All models developed in this thesis did not have their hyperparameters tuned with testing data.

The accuracy metric, or metrics, used to evaluate the models should be aligned with the goal of the model. The area under the curve (AUC) of the receiver operating characteristics is current and commonly used as the main metric when evaluating clinical models. In Chapter 5, the AUC stayed relatively stable when the prevalence of positive samples decreased. In that case, another metric, which is prevalence dependent, such as the Brier Score, was impacted by the prevalence changes. In this scenario, if the calibration of the model is important, analysing only the AUC would not show the prevalence change over time. Considering a more imbalanced scenario, such as the one presented in Chapter 8 where I aimed to support the diagnosis of a rare disease (with only a small fraction of positive samples), the proposed diagnostic support model had a relatively high AUC. However, despite the high AUC, the model was not specifically adjusted to have a high precision or recall. The threshold should have been adjusted to optimize a specific metric, trying to recover as many diseased patients as possible or trying to minimize the false positives. This illustrates that used metrics and thresholds should always be tailored in alignment with the desired use of the models and that metrics should be informative enough to ensure high accuracy when using the model in its intended use.

Final remarks and future directions

The creation of ML models depends intrinsically on data. Although cardiologists are becoming more familiar with AI, I understand that their focus should not be on learning AI theories but on how to improve the raw material used for the model’s creation. Although interoperability frameworks are becoming more common, there is no unique way of collecting and storing data, which makes it difficult to develop models using combined data from multiple centres and sources. Also, patient’s data is not constantly checked for consistency and quality,

even though this is an important step. Well-structured and validated procedures, as well as automatized data quality checks, should be considered by medical centres that aim to support AI researchers to create more optimized models.

The creation of ML models requires multiple steps and decisions need to be made when implementing them: how the models are optimized and validated, the pre-processing steps, and the metrics to be evaluated. The physicians should be involved in most of the steps and decisions, from model conception to validation, as they are the experts in the field where the models will be used. All the steps must be aligned with the physicians' needs to provide the best practice to patients. Even the definition of the target variable or features to be included could be very complex decisions. In Chapter 6, even though we focused on two outcomes, we evaluated prediction models for five different outcomes.

On occasions, it might not be simple to use the available data. In Chapter 7, for instance, the test bolus information is not easily accessible. When the data collection started for a previous study, the physicians accessed the files, one by one, and took notes of the test bolus information. Extracting information manually, however, is time-consuming and prone to errors when considering large amounts of data. For that reason, we implemented tools to automatically measure some information in the CT and applied reverse engineering to extract information from some of the files to promptly access the test bolus data.

Regarding the clinical use of the developed ML models, they should be used with caution. It is shown in this thesis that the models must be validated in many ways as they might have different sources of bias and flaws. Also, although ML models are becoming more common in clinical practice, the models themselves do not determine a diagnosis or predict the mortality of a patient. They estimate the likelihood of a particular outcome based on the data used for training the model. The models are proposed to be used as a decision-support tool to optimize physicians' (and patients') decisions and/or provide potentially new insights on a specific condition.

Evidently, the use of ML in cardiology is in a developing phase. Boundary conditions with respect to the quality of data and specific tools to be used are still in the investigation phase. Despite encouraging results reported, questions remain

not only on the quality of software development but also on the experts required to take care of the appropriate handling of tools. Although multidisciplinarity should be highly appreciated, specially trained engineers should be leading in developing tools. In addition, with the clinician's support, they must play a leading role in defining the quality of data input. Despite the current popularity of ML for advancing medical decision-making or highlighting previously underexposed associations, the quality of this technology application should be beyond doubt.

In this thesis, I explored the use of machine learning models for various topics in cardiology. In many of the chapters, although we used data from different sources, I did not focus on the use of multi-dimension raw data in most of the studies. The features were mainly collected manually by physicians/technicians (as scores or measurements) and presented as a small set of numeric values for the models to be trained. With DL approaches, like the one used in Chapter 8, multi-dimensional data (from ECGs or CTs) can be explored to let the models learn important features by themselves. Techniques like finetuning and distributed learning were applied to deal with data sharing policies and limited amounts of data for training the models. Model interpretation techniques and validation approaches, such as internal, external and temporal, were also explored. With these remarks, the work here presented also demonstrates the importance of continuous and proper evaluation of the models and data used.

References

1. Anaya-Isaza A, Mera-Jiménez L, Zequera-Diaz M. An overview of deep learning in medical imaging. Informatics Med Unlocked. 2021 Jan 1;26:100723.
2. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ [Internet]. 2020 Mar 25 [cited 2022 Jul 10];368. Available from: https://www.bmj.com/content/368/bmj.m689
3. Puvimanasinghe JPA, Steyerberg EW, Takkenberg JJM, Eijkemans MJC, Van Herwerden LA, Bogers AJJC, et al. Prognosis After Aortic Valve Replacement With a Bioprosthesis. Circulation [Internet]. 2001 Mar 20 [cited 2022 Sep 11];103(11):1535–41. Available from: https://www.ahajournals.org/doi/abs/10.1161/01.CIR.103.11.1535
4. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol [Internet]. 2018 Jun 12 [cited 2022 Jan 13];71(23):2668–79. Available from: https://pubmed.ncbi.nlm.nih.gov/29880128/
5. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial Intelligence in Cardiology: Present and Future. Mayo Clin Proc [Internet]. 2020 May 1 [cited 2022 Jan 13];95(5):1015–39. Available from: https://pubmed.ncbi.nlm.nih.gov/32370835/
6. Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Health Care. Ann Thorac Surg [Internet]. 2020 May 1 [cited 2022 Jan 13];109(5):1323–9. Available from: https://pubmed.ncbi.nlm.nih.gov/31706869/
7. Dudchenko A, Ganzinger M, Kopanitsa G. Machine Learning Algorithms in Cardiology Domain: A Systematic Review. Open Bioinforma J [Internet]. 2020 Apr 23 [cited 2022 Jan 4];13(1):25–40. Available from: https://openbioinformaticsjournal.com
8. Ongsulee P. Artificial intelligence, machine learning and deep learning. Int Conf ICT Knowl Eng. 2018;1–6.
9. Berggren K, Xia Q, Likharev KK, Strukov DB, Jiang H, Mikolajick T, et al. Roadmap on emerging hardware and technology for machine learning. Nanotechnology [Internet]. 2020 Oct 19 [cited 2022 Feb 23];32(1):012002. Available from: https://iopscience.iop.org/article/10.1088/1361-6528/aba70f
10. Andrew Ng Launches A Campaign For Data-Centric AI [Internet]. [cited 2022 Feb 23]. Available from: https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=4f1d8f1a74f5
11. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data [Internet]. 2018 Dec 1 [cited 2022 Feb 23];5(1):1–18. Available from: https://link.springer.com/articles/10.1186/s40537-017-0110-7
12. Kaur K, Rani R. Managing Data in Healthcare Information Systems: Many Models, One Solution. Computer (Long Beach Calif). 2015 Mar 1;48(3):52–9.
13. Mazumder M, Banbury C, Yao X, Karlaš B, Rojas WG, Diamos S, et al. DataPerf: Benchmarks for Data-Centric AI Development. arXiv Prepr arXiv220710062. 2022;

Summary

This thesis presented multiple strategies to train and validate machine learning-based prognostic, diagnostic, and decision support models in the field of cardiology. Prediction models are commonly used to support the decision of doctors and patients, however, they must be properly evaluated to avoid over-optimistic accuracies. For the evaluation of prediction models, depending on the desired use of the model, it is important to use an external population for validation. Also, it has been suggested that changes in procedures, population, and patient selection have important implications for models. The thesis explores multiple ways to assess and train outcome prediction models in **Chapters 2-6** and diagnostic and decision-support models in **Chapters 7-8** while dealing with several challenges.

Patients might have unfavourable outcomes after a Transcatheter Aortic Valve Implantation (TAVI) procedure, such as no improvement of symptoms or mortality. In **Chapter 2**, were investigated machine learning models in the prediction of 1-year mortality and improvement of symptoms after a TAVI procedure using screening and laboratory data. The accuracy of the different types of prediction models, such as linear and non-linear models, was similar when predicting mortality. Also, the accuracy of the models slightly improved when laboratory and screening data are combined. However, the accuracy of the models trained to predict the improvement of symptoms was rather low, independent of the features used.

In **Chapter 3**, I performed an external validation and finetuning of neural network-based TAVI mortality prediction models with data from two medical centres. I found that finetuning improved the overall accuracy of the models, especially for the centre that had the lowest number of patients. This study reassured the idea that combining data from multiple centres can potentially improve the models' accuracy.

Besides finetuning, there are other ways to train models without sharing data. Therefore, in **Chapter 4,** local and distributed approaches were explored to train

neural networks and tree-based algorithms for predicting mortality after TAVI. The models were trained in a distributed way and also trained locally followed by a model combination technique. Both centre-specific models had their accuracy improved with the proposed approaches. The larger centre's models had higher accuracy when using the stacking approach while the smaller centre's models had higher accuracy when training with the distributed approach.

Chapter 5 shows how the accuracy of TAVI mortality prediction models changed over the years because of changes in the procedure, patient selection and population. I evaluated the performance and stability of models trained once with only the oldest data available and compared with models being re-trained repeatedly over time including more recent data. The models that were re-trained repeatedly over time had improved stability compared to the model trained only once. The stability of the re-trained models mainly improved because of the decrease in patient mortality through the years, which was not considered in the model trained only once.

The study in **Chapter 6** focused on the outcome prediction after a thoracoscopic procedure for patients with atrial fibrillation. The baseline prediction model was created including all available features and the proposed model was based on features automatically selected, which improved the AUC significantly. The performed subgroup analysis showed reasonably high accuracy for younger patients.

In **Chapter 7** the focus shifts to the prediction of insufficient contrast enhancement on Coronary CT Angiography. The prediction models achieved higher accuracies when including test-bolus variables, which are not commonly used to adjust contrast delivery protocols. In addition, the test bolus' peak height was found as the feature that impacts the predictions the most, reinforcing that test bolus variables can be used to further improve patient-specific contrast delivery protocols.

Finally, in **Chapter 8**, I presented a diagnostic support model for the identification of patients with a rare genetic disease (PLN mutation) based on ECGs. The proposed approach, using a pre-trained model for sex prediction, outperformed the accuracy of models trained from scratch and previous studies.

This higher accuracy was observed with both balanced and imbalanced PLN-control ratio scenarios for training and testing.

Nederlandse samenvatting

In dit proefschrift worden meerdere strategieën gepresenteerd voor het trainen en valideren van op machine learning gebaseerde prognostische, diagnostische en beslissingsondersteunende modellen op het gebied van cardiologie. Voorspellingsmodellen worden vaak gebruikt om de beslissing van artsen en patiënten te ondersteunen, maar ze moeten goed geëvalueerd worden om te optimistische uitkomsten te vermijden. Voor de evaluatie van voorspellingsmodellen is het, afhankelijk van het gewenste doel van het model, van belang een externe populatie voor de validatie te gebruiken. Ook wordt aangetoond dat veranderingen in procedures, populatie en patiëntenselectie belangrijke implicaties hebben voor modellen. Het proefschrift verkent meerdere manieren om uitkomstvoorspellende modellen te beoordelen en te trainen in de **Hoofdstukken 2-6** en diagnostische en beslissingsondersteunende modellen in de **Hoofdstukken 7-8**, onderhevig verschillende uitdagingen.

Patiënten kunnen onwenselijke uitkomsten hebben na een Transcatheter Aortic Valve Implantation (TAVI) procedure, bijvoorbeeld aanhoudende symptomen na de TAVI procedure of sterfte. In **Hoofdstuk** 2, onderzochten we machine learning modellen in de voorspelling van 1-jaars mortaliteit en verbetering van symptomen na een TAVI procedure met behulp van screening en laboratorium gegevens. De nauwkeurigheid van de verschillende soorten voorspellingsmodellen, zoals lineaire en niet-lineaire modellen, was vergelijkbaar bij het voorspellen van mortaliteit. De nauwkeurigheid van de modellen licht wanneer laboratorium- en screeningsgegevens werden gecombineerd. Echter, de nauwkeurigheid van de modellen getraind om de afname van symptomen te voorspellen was vrij laag, onafhankelijk van de gebruikte kenmerken.

In **Hoofdstuk 3** voerde Ik een externe validatie en ‘finetuning’ uit van neurale netwerk-gebaseerde TAVI mortaliteitsvoorspellingsmodellen met gegevens van twee medische centra. Ik vond dat ‘finetuning’ de algemene nauwkeurigheid van de modellen verbeterde, vooral voor het centrum dat het laagste aantal patiënten

had. Deze studie bevestigt het idee dat het combineren van gegevens van meerdere centra de nauwkeurigheid van de modellen kan verbeteren.

Naast 'finetuning' zijn er ook andere manieren om modellen te trainen zonder gegevens te delen. Daarom werden in **Hoofdstuk 4**, lokale en gedistribueerde benaderingen

onderzocht om 'neural networks' en 'tree-based' algoritmen te trainen voor het voorspellen van mortaliteit na TAVI. De modellen werden op een gedistribueerde manier getraind en ook lokaal getraind, gevolgd door een model combinatie techniek. Beide centrum-specifieke modellen verbeterde de nauwkeurigheid door de voorgestelde toe te passe. De modellen van de grotere centra hadden een hogere nauwkeurigheid bij gebruik van de stapelingsbenadering, terwijl de modellen van de kleinere centra een hogere nauwkeurigheid hadden bij training met de gedistribueerde benadering.

Hoofdstuk 5 laat zien hoe de nauwkeurigheid van TAVI mortaliteitsvoorspellingsmodellen door de jaren heen veranderde door veranderingen in de procedure, patiëntenselectie en populatie. Ik evalueerden de prestaties en stabiliteit van modellen die eenmaal waren getraind met alleen langst bewaarde de beschikbare gegevens en vergeleek die met modellen die in de loop van de tijd herhaaldelijk opnieuw werden getraind, inclusief recentere gegevens. De modellen die in de loop van de tijd herhaaldelijk opnieuw werden getraind, hadden een betere stabiliteit dan het model dat slechts eenmaal was getraind. De stabiliteit van de opnieuw getrainde modellen verbeterde voornamelijk door de afname in patiëntensterfte door de jaren heen, waarmee geen rekening werd gehouden met het model dat slechts eenmaal was getraind.

De studie in **Hoofdstuk 6** richtte zich op de uitkomstvoorspelling na een thoracoscopische procedure voor patiënten met atriumfibrilleren. Het baseline voorspellingsmodel werd gemaakt inclusief alle beschikbare kenmerken en het voorgestelde model was gebaseerd op automatisch geselecteerde kenmerken, waardoor de AUC significant verbeterde. De uitgevoerde subgroep analyse toonde een redelijk hoge nauwkeurigheid voor jongere patiënten.

In **Hoofdstuk 7** ligt de focus naar de voorspelling van onvoldoende contrastversterking op coronaire CT-angiografie. De voorspellingsmodellen

bereikten een hogere nauwkeurigheid wanneer test-bolus variabelen werden meegenomen, die niet algemeen gebruikt worden om contrast toedieningsprotocollen aan te passen. Bovendien bleek de piekhoogte van de testbolus het meest van invloed te zijn op de voorspellingen, wat versterkt dat testbolusvariabelen gebruikt kunnen worden om patiëntspecifieke contrasttoedieningsprotocollen verder te verbeteren.

Tenslotte presenteerde Ik in **Hoofdstuk 8** een diagnostisch ondersteuningsmodel voor de identificatie van patiënten met een zeldzame genetische ziekte (PLN mutatie) op basis van het electrocardiogram. De voorgestelde aanpak, gebruikmakend van een vooraf getraind model voor geslachtsvoorspelling, overtrof de nauwkeurigheid van modellen getraind vanaf nul en eerdere studies. Deze hogere nauwkeurigheid werd waargenomen met zowel gebalanceerde als onevenwichtige PLN-controle ratio scenario's voor training en testen.

Abbreviations

AF	Atrial Fibrillation
AI	Artificial Intelligence
AMC	Academic Medical Center
ARB	Angiotensin Receptor Blockers
AS	Aortic Stenosis
AUC	Area Under the Curve
AUPRC	Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic
BMI	Body Mass Index
BS	Brier Score
BSA	Body Surface Area
CATB	CatBoost
CI	Confidence Interval
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration
CM	Contrast Material
CNN	Convolutional Neural Network
CO	Cardiac Output
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
CTCA	Computed Tomography Coronary Angiography
CV	Cross-Validation
CWT	Cyclical Weight Transfer
CZE	Catharina Ziekenhuis
DL	Deep Learning
DM	Diabetes Mellitus
ECG	Electrocardiogram
FEV1	Forced Expiratory Volume in One Second
FVC	Forced Vital Capacity
GBDT	Gradient Boosting Decision Tree
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GTB	Gradient Tree Boosting
HR	Heart Rate
HS-troponin	High Sensitive Troponin

HU	Hounsfield Units
ICD	Implantable Cardioverter Defibrillator
IDR	Iodine Delivery Rate
kV	Kilovolt
LA	Left Atrium
LASSO	Least Absolute Shrinkage and Selection Operator
LAVI	Left Atrial Volume Index
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-layer Perceptron
NT-proBNP	N-terminal pro-b-type Natriuretic Peptide
NYHA	New York Heart Association
PLN	Phospholamban
RF	Random Forest
RFC	Random Forest Classifier
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
RSPV	Right Superior Pulmonary Vein
SA	Surgical Ablation
SBP	Systolic Blood Pressure
SHAP	Shapley Additive exPlanations
SR	Sinus Rhythm
STS	Society of Thoracic Surgery
SVM	Support Vector Machines
TAVI	Transcatheter Aortic Valve Implantation
TIL	Total Iodine Load
TTE	Transthoracic Echocardiography
X-ECG	Exercise Testing ECG
XGB	eXtreme Gradient Boosting

Portfolio

Name PhD student: Ricardo Ricci Lopes

PhD period: April 2018 – March 2022

Name PhD supervisor: prof. dr. H.A. Marquering and prof. dr. B.A.J.M. de Mol

PhD training	**Year**	**ECTS**
General courses		
World of Science	2018	0.7
Practical Biostatistics	2018	1.1
E-science	2018	0.6
Second-generation p-values	2019	0.1
Unix	2020	0.5
Project Management	2020	0.6
Writing a Scientific Paper	2021	1.0
Didactical Skill	2022	0.4
Specific courses		
Advanced topics in Image Processing, UNICAMP	2016	2.1
Advanced topics in Computer Engineering, UNICAMP	2016	2.1
ECG Assessment, St George's University of London	2019	0.1
AI for Medical Diagnosis, Coursera	2020	0.7
International School on Deep Learning	2021	1.4
AI in Healthcare, Coursera	2022	2.0
Seminars, workshops and master classes		
Cardiovascular Engineering Meeting	2018-2022	4.0
Machine Learning Meeting (BMEP)	2018-2020	1.0
Journal Club – Machine Learning (BMEP)	2019	1.0

Symposium on Advances in Deep Learning	2019	0.2
Einstein Symposium on Cardiology	2020	0.3
Artificial Intelligence in Cardiology	2020	0.1
Amsterdam Medical Data Science	2019-2020	1.0
Presentations		
Machine Learning in Medical Imaging - iQC	2019	0.5
(Inter)national conferences		
World Summit AI, Amsterdam	2018	0.5
Medical Imaging Symposium for PhD Students - MISP, Rotterdam	2018	0.25
Medical Imaging with Deep Learning - MIDL, Amsterdam	2018	0.75
Medical Imaging with Deep Learning - MIDL, London	2019	0.75
Machine Learning in Medical Imaging - iQC, Amsterdam	2019	0.25
Computer Based Medical Systems - CBMS, online	2020	0.75

Teaching	**Year**	**ECTS**
Lecturing		
Information in medical images (MIK-BAM3.3)	2018	0.75
Advanced medical imaging processing (MIK-MAM10)	2018-2020	2.25
Supervising		
Master Internship - Tristan Chedeville	2019	0.5
Master Thesis - Casper van der Kerk	2018-2019	1.0
Bachelor Honours Program - Koen Kwakkenbos	2020-2021	1.0
Master Internship - Paul Fournier-Delouvée	2021	0.5
Master Thesis - Jussi Boersma	2021	0.5

Other

Reviewer European Radiology, LatinX in AI Research Workshop at NeurIPS, BMC Medical Informatics and Decision Making, Frontiers in Cardiovascular Medicine	2018-2022	1.0

List of publications

1. **Lopes RR**, van Mourik MS, Schaft EV, Ramos LA, Baan Jr J, Vendrik J, de Mol BA, Vis MM, Marquering HA. Value of machine learning in predicting TAVI outcomes. Netherlands Heart Journal. 2019 Sep;27(9):443-50.
2. **Lopes RR**, Mamprin M, Zelis JM, Tonino PA, van Mourik MS, Vis MM, Zinger S, de Mol BA, de With PH, Marquering HA. Inter-center cross-validation and finetuning without patient data sharing for predicting transcatheter aortic valve implantation outcome. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) 2020 Jul 28 (pp. 591-596). IEEE.
3. Baalman SW, Schroevers FE, Oakley AJ, Brouwer TF, van der Stuijt W, Bleijendaal H, Ramos LA, **Lopes RR**, Marquering HA, Knops RE, de Groot JR. A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples. International Journal of Cardiology. 2020 Oct 1;316:130-6.
4. Bleijendaal H, Ramos LA, **Lopes RR**, Verstraelen TE, Baalman SW, Pool MDO, Tjong FV, Melgarejo-Meseguer FM, Gimeno-Blanes FJ, Gimeno-Blanes JR, Amin AS, Winter MM, Marquering HA, Kok WEM, Zwinderman AH, Wilde AAM, Pinto YM. Computer versus cardiologist: is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p. Arg14del mutation on the electrocardiogram?. Heart Rhythm. 2021 Jan 1;18(1):79-87.
5. Mamprin M, **Lopes RR**, Zelis JM, Tonino PA, van Mourik MS, Vis MM, Zinger S, de Mol BA. Machine learning for predicting mortality in transcatheter aortic valve implantation: an inter-center cross validation study. Journal of Cardiovascular Development and Disease. 2021 Jun;8(6):65.

6. **Lopes RR**, Bleijendaal H, Ramos LA, Verstraelen TE, Amin AS, Wilde AAM, Pinto YM, de Mol BA, Marquering HA. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: An application to phospholamban p. Arg14del mutation carriers. Computers in Biology and Medicine. 2021 Apr 1;131:104262.
7. Baalman SW, **Lopes RR**, Ramos LA, Neefs J, Driessen AH, van Boven WP, de Mol BA, Marquering HA, de Groot JR. Prediction of atrial fibrillation recurrence after thoracoscopic surgical ablation using machine learning techniques. Diagnostics. 2021 Oct;11(10):1787.
8. van den Boogert TP, **Lopes RR**, Lobe NH, Verwest TA, Stoker J, Henriques JP, Marquering HA, Planken RN. Patient-tailored Contrast Delivery Protocols for Computed Tomography Coronary Angiography: Lower Contrast Dose and Better Image Quality. Journal of Thoracic Imaging. 2021 Nov 19;36(6):353-9.
9. **Lopes RR**, Mamprin M, Zelis J, Tonino PA, van Mourik M, Vis MM, Zinger S, de Mol BA, Marquering HA. Local and distributed machine learning for inter-hospital data utilization: an application for TAVI outcome prediction. Frontiers in Cardiovascular Medicine. 2021 Nov 12:1559.
10. **Lopes RR**, van den Boogert TP, Lobe NH, Verwest TA, Henriques JP, Marquering HA, Planken RN. Machine learning-based prediction of insufficient contrast enhancement in coronary computed tomography angiography. European Radiology. 2022 Jun 16:1-0.
11. Yordanov TR, **Lopes RR**, Ravelli AC, Vis M, Houterman S, Marquering H, Abu-Hanna A. An integrated approach to geographic validation helped scrutinize prediction model performance and its variability. Journal of Clinical Epidemiology. 2023 Feb 22.

Under review

1. **Lopes RR**, Yordanov TTR, Ravelli AACJ, Houterman S, Vis MM, de Mol BA, Marquering HA, Abu-Hanna A. Temporal validation of machine learning-based mortality prediction models for Transcatheter Aortic Valve Implantation (TAVI) using the Dutch Registry (NHR).

2. Terreros NA, Stolp J, Bruggeman AAE, Swijnenburg ISJ, **Lopes RR**, van Meenen LCC, Groot AED, Kappelhof M, Coutinho JM, Roos YBWEM, Emmer BJ, Beenen L, Dippel DWJ, van Zwam WH, van Bavel E, Marquering HA, Majoie CBLM. Thrombus Imaging Characteristics To Predict Early Recanalization In Transferred Patients With Large Vessel Occlusion Stroke.

Contributing authors

A Abu-Hanna
AS Amin
SW Baalman
J Baan Jr
H Bleijendaal
JR de Groot
AJM de Mol
BA de Mol
PH de With
AH Driessen
JP Henriques
S Houterman
WEM Kok
NH Lobe
M Mamprin
HA Marquering
J Neefs
YM Pinto
RN Planken
LA Ramos
AACJ Ravelli
EV Schaft
PA Tonino
WP van Boven
TP van den Boogert
MS van Mourik
J Vendrik
TE Verstraelen
TA Verwest
MM Vis
AAM Wilde
MM Winter
TTR Yordanov
JM Zelis
S Zinger
AH Zwinderman

Acknowledgements

At times during these 4 years in Amsterdam, it felt like it would never end. However, looking back now, it feels like it went by extremely fast. Despite all the challenges, I made it. Of course, I can't say that I have done it by myself only. I would like to express my gratitude not only for this achievement, but also for the process I went through.

Henk, first of all, thanks for the opportunity and for believing in me. I can't express how grateful I am for having you as a supervisor and colleague. Even though it felt like your comments on my manuscripts would never end, I must say that I've learned a lot from you. Thank you for supporting me when I needed it (more than a couple of times, I must say) and for supporting my decisions. I know that I can always count on you and I wish you all the best.

Bas, our conversations were always inspiring. Having someone like you, really excited about the work I was developing, only inspired me to work even harder. I believe we share a dream, and with your support, I'm happy to be taking steps towards achieving it.

Marije, Matthan, Nils, Yigal, Joris and Ameen. So many fruitful meetings and collaborations. Even with your busy schedules, you were always available to help me. I'm really happy for having the chance of working with you all. I would also like to express my gratitude to Martijn, Thomas, Sarah, and Hidde for their collaboration. I'm really happy with all the research we did and it was really nice to collaborate with you all. I wish you success with your bright career.

Manon, Henkie, Nerea, Praneeta, Marcela and Marco, you made my life much easier in the Netherlands. I cannot count the number of times you all have helped me. Besides that, it was amazing to you have around on so many days at the AMC, dinners, BBQ, pubs and parks.

Eva, Bart, Haryadi, Marit, Riaan, Renan, Bob, Nils, Lúcio, Tseko, and so many others that I had the chance to meet, work with, play, study and enjoy a cup of coffee/beer. Thank you!

I would like to thank my paranymphs, Lucas and Valenzuela. Lucas, in the end, I only got here because of you. From the moment you came to meet me at the airport on my first day in Amsterdam, it felt like we had been friends for a long time. Valenzuela, it was great to have an old friend coming to visit me but I never expected you to stay that long. Jokes apart, it was amazing to have you around. I feel really lucky for having the chance of working with you guys.

I can't say how thankful I'm for the support of my entire family but specially my parents, Cesar and Eleusa. Without your support, I would never have achieved what I achieved. Your support was one of the most important things and no words can describe how thankful I'm. I can only imagine how painful it was to see us leaving but we only did it because we knew we had you to come back. I love you always. Speaking of coming back, I cannot forget to mention my annoying sisters Ana e Carô. I would say that your calls were always enlightening my day (not only because of João and Matheus). I love you all.

Ju, I would never have made it without you by my side. Always supporting me, you had the most difficult task to hold me back and push me when needed. This achievement cannot be attributed solely to myself, as I always had you by my side. We knew it wouldn't be easy and, finally, we made it. I cannot deny that the sleepless nights were tough, but you definitely made them more bearable. You make me a better person, and I love you deeply!

About the author

Ricardo Ricci Lopes was born on the 15th of June in 1990 in Batatais, Brazil. He obtained his bachelor's degree in information systems and a master's degree in pattern recognition and computer vision from the Sao Paulo State University (UNESP). After finishing his master's, in 2015, he worked on developing machine learning models for image quality analysis at Eldorado Research Institute while pursuing an MBA in project management. Afterwards, he became the manager of the machine learning course of Udacity Brazil. In 2018, he moved to the Netherlands to pursue his PhD at the Amsterdam UMC (University of Amsterdam) in the biomedical engineering and physics department, where he researched the application of machine learning techniques in cardiology. He continued his career as a research scientist at Philips, Eindhoven.