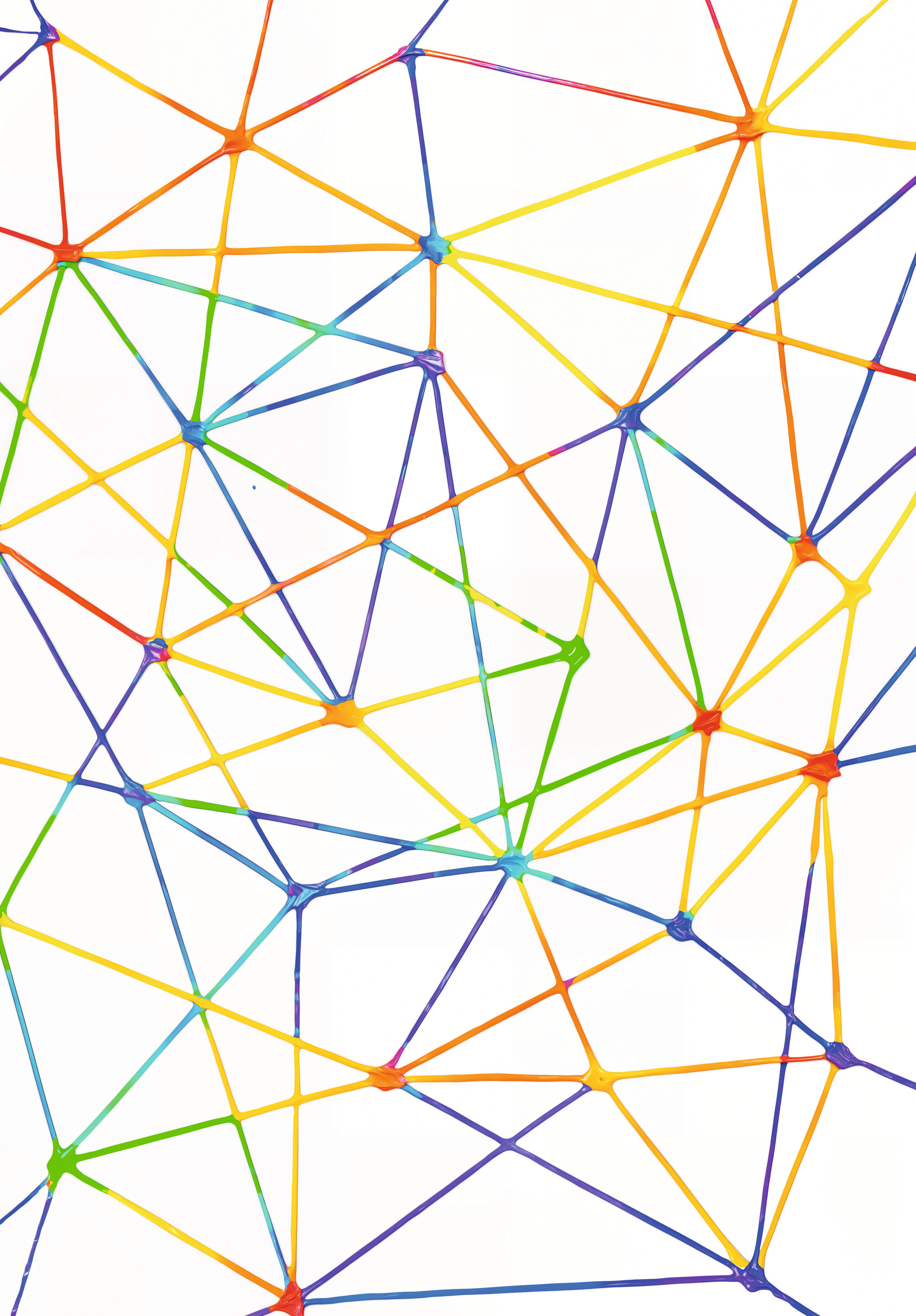

1

Introduction

Artificial intelligence (AI) has made extraordinary progress in the last couple of years, defeating top-ranking Go players [1], generating fantastical images [2], and writing grammatically flawless texts [3].[1] For the healthcare field, AI has the potential to improve patient outcomes through personalized predictions, lower healthcare costs, and reduce the administrative burden for healthcare professionals through automation and decision support [4]. In response to these big promises, recent years have seen a surge in healthcare AI funding [5] with clinical AI publications at an all-time high [6]. Yet, despite its potential, only a handful of AI algorithms have made it into healthcare practice [7-9]. Moreover, several safety concerns have been raised regarding the reliability of these algorithms when deployed in real-world settings [10, 11]. This thesis therefore focuses on the responsible development and validation of AI for healthcare.

1.1 PREDICTION ALGORITHMS

A clinical prediction algorithm aims to predict a usually binary outcome (e.g., hospital discharge or admission, or health outcomes such as cardiac infarction) by combining a number of characteristics [12]. In this thesis, prediction is used for prognostic purposes. This means that we predict the likelihood or risk of an event occurring in the future. Prediction algorithms can be developed using both classical statistics and AI techniques. Both types of techniques are discussed in more detail below.

1.1.1 Classical statistics techniques

There is a long history of developing prediction algorithms in the fields of epidemiology and medical statistics (referred to here as ‘classical statistics’). Examples of popular classical statistics algorithms are the Framingham risk score, predicting 10 year cardiovascular risk [13], and the 4C Mortality Score for risk stratification of COVID-19 patients [14]. Prediction algorithms derived with classical statistics methods usually rely on regression techniques, like logistic regression. A special class of methods called survival analysis may be used when the time between a baseline point (e.g., start of treatment) and the event of interest is important (e.g., death). A popular survival analysis technique is the Cox Proportional-Hazards model [12]. For situations with competing risks,

[1] See the chapter illustrations throughout this thesis and the summary for examples of image and text generation by Stable Diffusion and GPT3.

the Fine and Gray model is commonly used, while cause-specific Cox models may be preferred [15, 16].

1.1.2 Artificial intelligence techniques

This thesis is concerned with the subfield of AI known as machine learning (Figure 1.1). Machine learning is a class of AI techniques that learns from data to improve its performance on a prediction task. Typically, we do not specify the full set of (interaction) terms for the algorithm to learn from, as might be needed for a classical statistics approach. As a result, machine learning algorithms provide us with a lot of flexibility to model nonlinear relationships [17]. Because machine learning algorithms can become highly complex and difficult to interpret, they are also referred to as black boxes [18, 19].

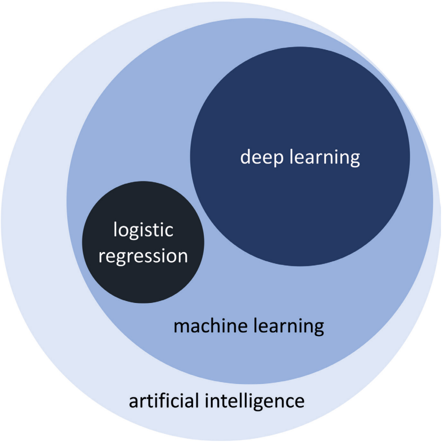

Figure 1.1 AI hierarchy and terminology

The most prominently featured machine learning algorithms in this thesis are tree-based algorithms (e.g., random forest and gradient boosting decision trees) and neural networks. A random forest combines the predictions of many independently built decision trees into one prediction (also called bagging) [20]. Gradient boosting decision trees are essentially a random forest that is optimized through gradient boosting [21]. Neural networks are comprised of several layers of nodes that are highly interconnected [22]. A neural network with one layer, one node and a sigmoid activation function is akin to a logistic regression (Figure 1.2). The depth created by the layers in the network is why this type of machine learning is also called deep learning (Figure 1.1).

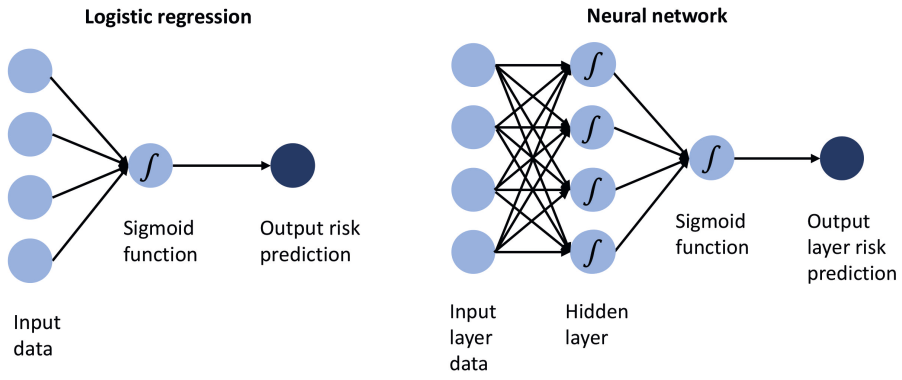

Figure 1.2 Schematic representation of a logistic regression and a neural network

1.2 TOWARDS IMPLEMENTATION: DEVELOPMENT AND VALIDATION

This thesis discusses AI prediction algorithms that are intended for implementation in healthcare practice. The development and validation steps are crucial to successful and safe implementation and are the focus of this thesis. The objective of algorithm development is to train an algorithm that most accurately predicts the outcome variable on new, unseen data. Part of the dataset is selected to train the algorithm. The leftover data (also holdout, validation, or test data) is used to test the algorithm's performance in the same setting as the train data, but with new examples not used during training. There are several designs to optimize this training procedure, like *k*-fold cross-validation and bootstrapping [23]. Algorithm development is followed by validation in which the performance of the prediction algorithm is tested on an external dataset [24]. Validation is needed to assess an algorithm's generalizability to other settings that are different from the development setting (for example in time or place).

1.2.1 Assessing algorithm performance

During algorithm development and validation, we need to quantify an algorithm's prediction performance. This can be done along various axes. Classification measures quantify the algorithm's ability to correctly classify patients at a specific decision threshold. For example, at a decision threshold of 50%, risk predictions higher than 50% are classified as 1 for experiencing the outcome and predictions lower than 50% as 0 for not experiencing it. Classification measures are based on the 2x2 confusion matrix (Table 1.1). Common classification measures are accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Discrimination measures such as the Area Under the Receiver Operating Characteristic curve (AUROC) are also based on the confusion matrix but evaluated for consecutive decision thresholds. They quantify the separation between low-risk and high-risk individuals [25] (Table 1.2). Calibration assesses the reliability of the risk predictions [25, 26], for example through a calibration curve (Table 1.2). Calibration is good when the proportion of the individuals receiving a certain risk score approximates that risk score. For example, within the group of patients receiving a 30% risk prediction of hospital admission, 30% of patients are in fact admitted to the hospital. Lastly, utility measures such as Net Benefit measure the number of true-positive classifications penalized for false-positive classifications [27-29] (Table 1.2). It provides a first indication of the clinical utility of the prediction algorithm with respect to current healthcare practice.

Table 1.1 Confusion matrix for classification measures at a specific decision threshold

Event Algorihm	True event	False events	
True predictions	True positive (TP)	False positive (FP)	Positive Predictive Value $\frac{TP}{TP+FP}$
False predictions	False negative (FN)	True negative (TN)	Negative Predictive Value $\frac{TN}{TN+FN}$
	Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{TN+FP}$	Accuracy $\frac{TP+TN}{TP+FP+TN+FN}$

Table 1.2 Classification, discrimination, calibration, and clinical utility evaluation measures

Evaluation measures	Definition
Classification	Quantifies an algorithm's ability to correctly classify cases at a specific decision threshold.
Accuracy; sensitivity; specificity; positive predictive value; negative predictive value	See Table 1.1.
Discrimination	Quantifies the separation between low-risk and high-risk individuals.
Area under the receiver operating characteristic curve (AUROC)	The receiver operating characteristic curve plots sensitivity as a function of 1-specificity for consecutive decision thresholds to classify a patient as high risk.
Area under the prediction recall curve (AUPRC)	The precision recall curve plots the positive predictive value (precision) as a function of sensitivity (recall).
Calibration	Quantifies the reliability of the risk predictions.
Calibration curve	A calibration curve plots the agreement between observed and predicted risks [25, 26]. The intercept relates to calibration-in-the-large and the correct estimation of overall baseline risk. An intercept below 0 indicates that the estimated risks are too high and above 0 too low. The slope measures whether risks are too extreme or modest. A slope below 1 implies too extreme risk estimates, above 1 too moderate.
Calibration loss	The calibration loss is calculated by ordering all cases by their risk estimate [30]. Cases 1-100 are put in the same bin. The percentage of true positives and the mean prediction are calculated. The absolute value of the difference between the true positive percentage and the mean prediction is the calibration error for this bin. Repeat this for cases 2-101, 3-102, etc. The calibration loss is the mean of all the binned calibration errors.
Clinical utility	Quantifies benefits and harms of an algorithm for clinical decision making.
Net Benefit	Net Benefit is a weighted sum of true positive (TP) and false positive (FP) predictions at a given decision threshold (t): $NB = (TP - \frac{t}{1-t} * FP)/N$ [27-29] Net Benefit can be plotted over a range of decision thresholds resulting in a decision curve.

1.3 CHALLENGES FOR RESPONSIBLE AI IN HEALTHCARE

Despite the promises of AI and machine learning for healthcare, we note a lack of systematic development and validation practices for clinical AI prediction algorithms: Guidance is incomplete or missing, jargon differs between the statistics and computer science fields [31], and external validation is rare [32]. This can have real-world ramifications, as illustrated by the widely used Epic sepsis algorithm that failed to detect 67% of patients with sepsis [11]. To create realistic expectations for AI in healthcare and facilitate responsible development, validation, and implementation, knowledge is required on the boundary conditions for a successful AI project. What situations benefit from the use of AI techniques and in what situations could it be disadvantageous or even harmful? What should be taken into consideration when applying AI prediction algorithms to healthcare practice?

1.4 AIMS

The overall aim of this thesis is to provide insight in the responsible development and validation of AI algorithms for outcome prediction in healthcare practice. To this end, this thesis has three specific research questions:

1. What are prime considerations in the development and validation of artificial intelligence prediction algorithms?

2. What are opportunities for classical statistics and artificial intelligence techniques for developing prediction algorithms?

3. What is the performance and potential of artificial intelligence prediction algorithms for clinical practice?

1.5 DATA

The data for the prediction algorithms in this thesis come from a variety of sources, allowing us to study the applicability of AI prediction algorithms in different healthcare settings (Table 1.3). Most use cases focus on the application of a prediction algorithm in a hospital environment. The clinical domains cover emergency medicine, primary care medicine, orthopedic surgery, oncology, and intensive care medicine. Most datasets are derived from hospital Electronic Health Records (EHR). All datasets consist of tabular (or structured) data fields. One study also uses unstructured text data. The acquired datasets generally consist of real-time data, allowing us to study the performance of AI prediction algorithms in real-world settings.

Table 1.3 Characteristics of the datasets used in this thesis

CH	Clinical domain	Data source	Data type	Data modality
4	Emergency medicine	NEED registry [33]	EHR	Tabular data
5	Primary care medicine	Controlled trials [34, 35]	Patient diaries	Tabular data
6	Arthroplasty surgery	LROI registry [36]	Questionnaires	Tabular data
7	Oncology	Stanford data lake	EHR	Tabular and text data
8	Intensive care medicine	LUMC data lake	EHR	Tabular data

1.6 OUTLINE

This thesis is divided into two parts.

Part I of this thesis is concerned with methods: the description of guidance and several methodological considerations for the development, validation, and implementation of clinical AI prediction algorithms. **Chapter 2** contains a scoping review that identifies actionable guidance for all stages of AI development. **Chapter 3** includes three short articles reflecting on validation methods. It describes i) the methodological differences between the statistics and computer science fields for algorithm evaluation, ii) how to (not) interpret the area under the receiver operating characteristic curve, and iii) the necessity to align one's validation goals with an algorithm's intended use.

Part II describes development and validation of prediction algorithms in five use cases. In **chapters 4, 5, and 6** classical statistics and AI techniques are compared. **Chapter 4** covers the development of several AI prediction algorithms for predicting hospital admission at the emergency department. **Chapter 5** compares different algorithms for predicting severe asthma exacerbations in asthma patients. In **chapter 6** several competing risk algorithms are developed to predict arthroplasty revision. **Chapter 7** assesses the value of multimodal data for predicting depression risk in oncology patients at the start of treatment. **Chapter 8** discusses the external validation and retraining of a proprietary algorithm predicting ICU readmission or death within seven days of discharge. In **chapter 9**, the main findings of this thesis and their implications for research and development are discussed.

REFERENCES

1. Silver, D., A. Huang, C.J. Maddison, et al., *Mastering the game of Go with deep neural networks and tree search.* Nature, 2016. **529**(7587): p. 484-489.
2. Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical text-conditional image generation with clip latents.* arXiv preprint arXiv:2204.06125, 2022.
3. Ouyang, L., J. Wu, X. Jiang, et al., *Training language models to follow instructions with human feedback.* arXiv preprint arXiv:2203.02155, 2022.
4. Yu, K.H. and I.S. Kohane, *Framing the challenges of artificial intelligence in medicine.* BMJ Qual Saf, 2019. **28**(3): p. 238-241.
5. Ben Leonard and Ruth Reader. Artificial intelligence was supposed to transform health care. It hasn't. POLITICO, 2022. Available from: https://www.politico.com/news/2022/08/15/artificial-intelligence-health-care-00051828.
6. Secinaro, S., D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, *The role of artificial intelligence in healthcare: a structured literature review.* BMC Medical Informatics and Decision Making, 2021. **21**(1): p. 125.
7. Smith, M., A. Sattler, G. Hong, and S. Lin, *From Code to Bedside: Implementing Artificial Intelligence Using Quality Improvement Methods.* J Gen Intern Med, 2021.
8. Vollmer, S., B.A. Mateen, G. Bohner, et al., *Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness.* Bmj-British Medical Journal, 2020. **368**.
9. Wiens, J., S. Saria, M. Sendak, et al., *Do no harm: a roadmap for responsible machine learning for health care.* Nat Med, 2019. **25**(9): p. 1337-1340.
10. Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations.* Science, 2019. **366**(6464): p. 447-453.
11. Wong, A., E. Otles, J.P. Donnelly, et al., *External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients.* JAMA Internal Medicine, 2021. **181**(8): p. 1065-1070.
12. Steyerberg, E.W., *Clinical Prediction Models,* ed. M. Gail, M.S. Jonathan, and B. Singer. 2009, Cham, Switzerland: Springer Nature.
13. Wilson, P.W., R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel, *Prediction of coronary heart disease using risk factor categories.* Circulation, 1998. **97**(18): p. 1837-47.
14. Knight, S.R., A. Ho, R. Pius, et al., *Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score.* BMJ, 2020. **370**: p. m3339.
15. Fine, J.P. and R.J. Gray, *A Proportional Hazards Model for the Subdistribution of a Competing Risk.* Journal of the American Statistical Association, 1999. **94**(446): p. 496-509.
16. Austin, P.C., E.W. Steyerberg, and H. Putter, *Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1.* Stat Med, 2021. **40**(19): p. 4200-4212.
17. Vapnik, V., *The nature of statistical learning theory.* 1999: Springer science & business media.

18. Röösli, E., S. Bozkurt, and T. Hernandez-Boussard, *Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model.* Scientific Data, 2022. **9**(1): p. 24.
19. Adadi, A. and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).* IEEE Access, 2018. **6**: p. 52138-52160.
20. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
21. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
22. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* The bulletin of mathematical biophysics, 1943. **5**(4): p. 115-133.
23. Steyerberg, E.W., *Validation in prediction research: the waste by data splitting.* Journal of Clinical Epidemiology, 2018. **103**: p. 131-133.
24. Moons, K.G., A.P. Kengne, D.E. Grobbee, et al., *Risk prediction models: II. External validation, model updating, and impact assessment.* Heart, 2012. **98**(9): p. 691-8.
25. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.
26. Van Calster, B., D.J. McLernon, M. van Smeden, et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
27. Vickers, A.J., B. Van Calster, and E.W. Steyerberg, *Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests.* BMJ, 2016. **352**: p. i6.
28. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models.* Med Decis Making, 2006. **26**(6): p. 565-74.
29. Vickers, A.J., B. van Calster, and E.W. Steyerberg, *A simple, step-by-step guide to interpreting decision curve analysis.* Diagnostic and Prognostic Research, 2019. **3**(1): p. 18.
30. Caruana, R. and A. Niculescu-Mizil. *Data mining in metric space: an empirical analysis of supervised learning performance criteria*. in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004.
31. Faes, L., D.A. Sim, M. van Smeden, U. Held, P.M. Bossuyt, and L.M. Bachmann, *Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?* Front Digit Health, 2022. **4**: p. 833912.
32. Wu, E., K. Wu, R. Daneshjou, D. Ouyang, D.E. Ho, and J. Zou, *How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals.* Nature Medicine, 2021. **27**(4): p. 582-584.
33. *Netherlands Emergency department Evaluation Database (NEED)*. Available from: https://www.stichting-need.nl.
34. Smith, A.D., J.O. Cowan, K.P. Brassett, G.P. Herbison, and D.R. Taylor, *Use of Exhaled Nitric Oxide Measurements to Guide Treatment in Chronic Asthma.* New England Journal of Medicine, 2005. **352**(21): p. 2163-2173.
35. Taylor, D.R., G.I. Town, G.P. Herbison, et al., *Asthma control during long term treatment with regular inhaled salbutamol and salmeterol.* Thorax, 1998. **53**(9): p. 744.
36. *Landelijke Registratie Orthopedische Interventies (LROI)*. Available from: https://www.lroi.nl/.