



Machine learning models in clinical practice for the prediction of postoperative complications after major abdominal surgery

Wessel T. Stam^{1,2,3} · Erik W. Ingwersen^{1,2,3} · Mahsoem Ali^{1,2} · Jorik T. Spijkerman⁴ · Geert Kazemier^{1,2} · Emma R. J. Bruns^{1,2} · Freek Daams^{1,2}

Received: 21 July 2022 / Accepted: 7 February 2023 / Published online: 25 February 2023
© The Author(s) 2023

Abstract

Complications after surgery have a major impact on short- and long-term outcomes, and decades of technological advancement have not yet led to the eradication of their risk. The accurate prediction of complications, recently enhanced by the development of machine learning algorithms, has the potential to completely reshape surgical patient management. In this paper, we reflect on multiple issues facing the implementation of machine learning, from the development to the actual implementation of machine learning models in daily clinical practice, providing suggestions on the use of machine learning models for predicting postoperative complications after major abdominal surgery.

Keywords Machine learning · Prediction · Postoperative complications

Introduction

Surgical patients are inevitably at risk of suffering postoperative complications, despite decades of scientific and technological advancement. The accurate prediction of individual outcomes has the potential to completely reshape the future of postoperative management. Such prediction would enable shared clinical decision-making and individual perioperative care and postoperative management.

In the last few years, a number of prediction models have been developed using machine learning (ML) models. These models offer the opportunity to develop a more individualized approach, allowing for data-driven individualized medicine [1–3]. However, clinical implementation and

acceptance are cumbersome, as it is often hampered by non-compliance with necessary guidelines.

To achieve the transparent, safe, and applicable implementation of ML models in the prediction of postoperative outcomes, we propose uniform selection, training, and guideline compliance.

Barriers and solutions

There is a large gap between promising and comprehensive research on the potential utility of artificial intelligence (AI) in the field of medicine and its actual implementation in daily clinical practice [4]. Several authors have tried to use ML models to optimize postoperative management by predicting postoperative complications. Unfortunately, many conclude that the implementation of these models is far from being clinically viable, even though most ML models achieve reasonable performance [5–15]. Cao et al. [5], Weller et al. [12], and Van den Bosch et al. [15] concluded that no practical implementation could be achieved for ML models due to the predictive value being too low to clinically implement. Grass et al. predicted surgical site infection in patients after colorectal surgery and initially found that ML outperformed conventional logistic regression. However, after external validation the practical applicability dropped due to low predictive performance [7].

✉ Freek Daams
f.daams@amsterdamumc.nl

¹ Department of Surgery, Amsterdam UMC Location Vrije Universiteit Amsterdam, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands

² Cancer Center Amsterdam, Cancer Treatment and Quality of Life, Amsterdam, The Netherlands

³ AGEM Amsterdam Gastroenterology, Endocrinology and Metabolism, Amsterdam, The Netherlands

⁴ Independent Consultant in Computational Intelligence, Amsterdam, The Netherlands

Table 1 summarizes the outcomes of the largest, most recent articles on postoperative complication prediction with ML in major abdominal surgery. The shift to clinical implementation will depend on five main improvement categories: technology, policy, medical and economic impact, transparency, and reporting [4, 16, 17].

Selection of a model

Model selection in conventional statistics is well defined due to strict requirements, as opposed to ML models, in which it typically depends on several factors. Even experienced data scientists have difficulty selecting an optimal model [18].

Model selection in ML, which involves determining the highest overoptimism-corrected performance in the metric and output of choice, is performed post hoc by exploring different models. Factors of consideration are the quality and size of the data, questions that must be answered, time available for running the model, and type of desired output. Depending on the hypothesis, different metrics can be used to measure this performance. For example, in the prediction of events with a relatively low incidence (e.g. anastomotic leakages after hemicolectomy), imbalanced data may occur. In such cases, using a single metric to score outcome such as accuracy is less effective, therefore, precision, recall, and the *f1*-score are more explanatory. Although it is advisable to use multiple metrics to measure performance more broadly, most studies published thus far on the use of AI for predicting complications did not adhere to this principle. The majority of such studies focused only on sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC) [19, 20].

The complexity of ML models may lead to challenges in understanding the involved mechanisms [21]. Many argue that it is highly inadvisable to rely on outcomes of ‘black-box’ systems in the decision-making process, as this can ignore the moral responsibilities of medical professionals. However, at the same time, we regularly prescribe medications, such as acetaminophen, without fully understanding the mechanism [22]. It is argued that, without pursuing the ‘explainability’ of AI tools, better outcomes for patients cannot be provided [23–25].

In addition, many claim that carrying out the decision-making process by a ‘black box’ system carries inherent dangers [26–28]. However, the importance of these dangers varies based on the ethical burden of the decisions that depend on it, thus suggesting that not knowing what a model is based on does not necessarily mean it should be seen as a danger [29]. For example, in the prediction of postoperative mortality, model explainability might be of more value than the prediction of postoperative delayed gastric emptying after distal gastrectomy. As Aristotle stated over two millennia ago, “the ability to verify results by empirical

means are more important than to explain the etiology of these results.” This is particularly important in a field in which knowledge of causality is often incomplete, as with postoperative sequelae [21].

Proving that a complication can be predicted while having the ability to reproduce these results might be more important than how this prediction is made. Therefore, a more complex and less explanatory model than a transparent, but simple model might be acceptable for predicting complications [30]. However, to maximize the model interpretability, one could use either individual conditional expectations (ICEs), local interpretable model-agnostic explanations (LIMEs), or Shapley additive explanations (SHAPs) [31]. These techniques aim to increase the comprehensibility of the rationale behind the model’s prediction by visualizing the contributing impact of different variables. This offers the possibility to approach the mechanisms within black-box systems in a way that empowers the clinician to trust the results produced by these models.

Training, validating, and testing

The generalizability of an ML model depends on the extent and quality of its training, validating, and testing. A very complex model for predicting postoperative mortality after pancreatic surgery might perform nearly perfectly during training but might not be able to properly predict the risk prospectively. This phenomenon is called overfitting, and it occurs when a model is incapable of capturing the relationship between the input variables and the target output values.

One way to estimate the extent of overfitting is through repeated cross-validation or preferentially via bootstrap resampling. This is where the entire modelling process is repeated in each bootstrap replicate. Quantifying the degree of overoptimism in the model’s performance also enables the observation of bias-corrected model performance estimates [20, 32]. Bootstrap resampling with preferentially 200 to 1000 bootstrap replications can provide stable and accurate overoptimism-corrected performance estimates, which has made it the gold standard for internal validation [33, 34]. In contrast, when there is a high error rate in both the training and testing data due to high bias and low variance, the model may be underfit. The balance between underfitting and overfitting is called the bias-variance trade-off [35]. This trade-off shows that when the complexity of a model increases, variance also increases, and bias falls.

Infrastructure and transparency

Adapting facility infrastructure to enable safe, real-time interaction between the patient file and ML models is both time-consuming and expensive [36]. The implementation of working models in other facilities can be equally difficult

Table 1 Description of recent large-scale studies on ML and complication predictions as well as key findings

First author (year)	Study design	Data source	Population	Prediction	Methods	Outcome
Cao (2019)	Retrospective	National database	N: 44,061 Surgical specialty: bariatric surgery Number of patients with complications: 1408 (3.2%)	30-day severe postoperative complications	Comparison of 29 ML models Internal and external validation	No achievement of practical applicability due to low performance
Chen (2018)	Retrospective	Institutional database	N: 13,399 Surgical specialty: colorectal surgery Number of patients with complications: 1680 (12.5%)	7-day postoperative bleeding	Overfitting prevention Comparison of an ML model to conventional LR Internal validation, no external validation Overfitting prevention	ML models outperformed conventional LR models. They potentially allow for targeted surveillance
Grass (2020)	Retrospective	National database	N: 2376 Surgical specialty: colorectal surgery Number of patients with complications: 108 (4.6%)	Surgical site infection	Comparison of an ML model to conventional LR Internal validation, no external validation	ML models outperformed conventional LR models. However external validation showed low performance. Institutional data were advised
Han (2020)	Retrospective	Institutional database	N: 1769 Surgical specialty: pancreaticoduodenectomy Number of patients with complications: 221 (12.5%)	Postoperative pancreatic fistula	Two ML models tested Internal validation, no external validation	ML models outperformed conventional LR models
Merath (2020)	Retrospective	National database	N: 15,657 Surgical specialty: liver (6012 [38.4%]), pancreatic (685 [4.4%]), and colorectal (8960 [57.2%]) Number of patients with complications: 6073 (38.8%)	30-day postoperative complications	Comparison of an ML model to conventional risk scores Internal validation, no external validation	ML models outperformed conventional LR models
Nudel (2021)	Retrospective	National database	N: 436,807 Surgical specialty: bariatric surgery Number of patients with complications: 3,068 (0.70%) developed anastomotic leakage, 2012 (0.46%) developed venous thromboembolism	30-day postoperative anastomotic leakage and venous thromboembolism	Comparison of two ML models to conventional LR Internal and external Overfitting prevention	ML models outperformed conventional LR models regarding anastomotic leakage LR models outperformed ML models regarding venous thromboembolism
Pera (2022)	Retrospective	National database	N: 3182 Surgical specialty: gastric surgery Number of patients with complications: 178 (5.6%)	90-day postoperative mortality	Comparison of three ML models to conventional LR Internal validation, no external validation Overfitting prevention	Good ML model performance, but due to no external validation and an imbalanced dataset, outcomes may not be extrapolatable
Shi (2012)	Retrospective	National database	N: 22,926 Surgical specialty: liver Number of patients with complications: 619 (2.7%)	In-hospital mortality	Comparison of two ML models to conventional LR Internal and external validation	ML models outperformed conventional LR models

Table 1 (continued)

First author (year)	Study design	Data source	Population	Prediction	Methods	Outcome
Van den Bosch (2022)	Retrospective	National database	N: 62,501 Surgical specialty: colorectal surgery Number of patients with complications: 1693 (2.7%)	30-day postoperative mortality	Comparison of four ML models Internal validation, no external validation Overfitting prevention	ML models outperformed conventional LR models. However, there was no achievement of practical applicability due to low performance
Weller (2018)	Retrospective	Institutional database	N: 9598 Surgical specialty: colorectal surgery	Postoperative complications	Comparison of four ML models Internal and external validation	There was no achievement of practical applicability due to low performance
Wise (2019)	Retrospective	National database	N: 101,721 Surgical specialty: bariatric surgery Number of patients with complications: 3,853 (3.8%)	30-day postoperative morbidity and mortality	Comparison of an ML model to conventional LR Internal validation, no external validation	LR models outperformed ML models

N number, ML machine learning, LR logistic regression

since models are often facility-specific [10, 12, 36, 37]. It is therefore of paramount importance that ML models be trained and validated within multiple healthcare facilities with adequate sample sizes to ensure generalizability as well as to prevent substantial harm to the patient [38–40]. In addition to such efforts, the use of a uniform classification system, or international consensus, is a prerequisite for ensuring clinical applicability and generalizability [37].

While abundant literature exists on the methodology and reporting quality of models using conventional statistics (i.e. logistic regression), there is increasing concern about the transparency of studies using ML models [17, 41]. Suboptimal transparency in model development makes ML models hard to interpret, which thus impedes their implementation. This is regarded as the main reason for the limited application of ML models in daily clinical practice [42, 43]. In addition to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement (TRIPOD), a protocol for ML-specific prediction was recently published to optimize the transparency of prediction models [17, 44]. Adherence to this statement will improve the interpretability, reproducibility, risk of bias assessment, and ultimately its applicability in clinical practice [45]. The completeness of this checklist is generally poor, as only 38.7% of the 152 articles published using ML models for prediction adhered to the TRIPOD items [16]. Model specification alongside model performance, which are both essential in transparency reports, were also rarely reported [16].

Data and interpretability

The large effect of the data quality on the final results is an often-mentioned pitfall in AI implementation and is called the Garbage-In-Garbage-Out principle [46]. Inaccurate data directly lead to unreliable results. The World Health Organization stated that proper data quality is multidimensional and should be accurate, available, complete, and valid [46–48]. This quality should always be accurate for ML purposes. The type of data used to feed the model depends on the moment a prediction must be made. Xue et al. [49] evaluated the utility of pre- and intraoperative data for predicting postoperative complications. They concluded that having a combination of pre- and intraoperative data resulted in slightly better performance than an analysis with only one of the two types.

Furthermore, a combination of structured and unstructured data is advisable. With unstructured data in particular, such as data from electronic health records and computed tomography or magnetic resonance imaging, ML models show their superiority [50]. It is therefore advisable to use a combination of both data types when possible to help formulate a risk score based on as many contributing static and modifiable risk factors as possible, allowing for early intervention. When risk scores are created, it is essential

that several experts in the field of surgery, nursery, or data science discuss the favorable cut-off values together. The previously discussed ICEs, LIMEs, and SHAPs can also contribute to the interpretability of the model's output.

Acceptance and ethical considerations

AI has great potential utility for healthcare professionals in supporting or augmenting clinical decision-making. Multiple studies suggest that these models will play a critical role in future surgical decision-making [51]. However, even when a model has been tested and validated correctly, its degree of acceptance by clinicians and patients can greatly affect its implementation.

Patients correlate AI with science fiction, drawing a fear of machines and computers taking over the making of decisions affecting human beings. The majority of patients thus prefer to receive a healthcare provider's supervision over AI [52]. The fear of clinicians being replaced by AI leads to mistrust of these models by patients. Similarly, in a study on the acceptance of AI amongst clinicians, only 25% of radiologists had confidence in the results of diagnoses made by AI algorithms [53].

Therefore, medical tools using AI should be used in an assistive manner as opposed to being ultimately responsible for the main decision-making [54]. The 'doctor in the loop' is responsible and this responsibility is classified into the following: accountability, liability, and culpability [54, 55]. To tackle this problem it is important to have proper patient education to reassure that the AI systems are not replacing the decision-making of the professional and are merely acting in a complementary manner to them.

Before the predictions of ML models can be used in daily clinical practice successfully, healthcare providers need to achieve trust in these techniques. Surgeons deciding to restore colonic continuity after resection of colon cancer based on an intraoperative ML model predicting a high risk of anastomotic leakage will want to rely on algorithms that have been validated at their own institute. A prospective simulation study in which the predictive performance of a model is tested in addition to the regular local care without affecting its course would enable the correct calibration of the ML model. A calibration curve shows whether or not the predicted chance matches the actual population-based chance of developing a postoperative complication. This approach would allow surgeons to attain trust in the effectiveness and predictive performance of ML models and use them in their clinical practice [56, 57].

The introduction of ML models has led to an unprecedented amount of ethical issues, and guidelines regarding these ethical considerations are still sparse [58]. Currently available frameworks for governance were discussed in a recently published review of the literature [59]. This study

included 21 guidelines for gold-standard societal values, such as sustainability, freedom, and fairness. Although these guidelines appeared to be insufficient when analyzed separately, it was stated that the ideal rules for ethical considerations should harmonize interests, offering benefit to clinicians, patients, and hospitals [59]. A governance model for the application of ML models in healthcare based on the abovementioned concept was developed recently [60]. In our opinion, it is of utmost importance to adhere to such governance models to ensure acceptance as well as ethical and legal appropriateness.

Correct prediction of postoperative complications using ML has the potential to dramatically improve the outcome of everyday clinical surgical care. However, their implications for patients should be considered before implementing prediction models in clinical practice. For example, predicting anastomotic leakage after colorectal surgery may lead to more stomas or earlier discharge when leakage is not expected to occur. This could also lead to collateral over- and under-treatment. This change in the paradigm of clinical practice must be accepted by all healthcare providers to ensure full benefit from these techniques. Therefore, it is advisable to obtain solid prospective validation from external sources at different centers with adequate sample sizes, all while adhering to the transparency and ethical guidelines to overcome potential distrust concerning ML among clinicians and patients.

Declarations

Conflict of interest All authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Machine learning for patient risk stratification. Standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med.* 2021;4(1):62.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
3. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA.* 2016;316(22):2368–9.

4. Wolff J, Pauling J, Keck A, Baumbach J. Success factors of artificial intelligence implementation in healthcare. *Front Digit Health*. 2021;3: 594971.
5. Cao Y, Fang X, Ottosson J, Näslund E, Stenberg E. A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. *J Clin Med*. 2019;8(5):668.
6. Chen D, Afzal N, Sohn S, Habermann EB, Naessens JM, Larson DW, et al. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. *Surgery*. 2018;164(6):1209–16.
7. Grass F, Storlie CB, Mathis KL, Bergquist JR, Asai S, Boughey JC, et al. Challenges of modeling outcomes for surgical infections: a word of caution. *Surg Infect*. 2020;22:523.
8. Han IW, Cho K, Ryu Y, Shin SH, Heo JS, Choi DW, et al. Risk prediction platform for pancreatic fistula after pancreatoduodenectomy using artificial intelligence. *World J Gastroenterol*. 2020;26(30):4453–64.
9. Merath K, Hyer JM, Mehta R, Farooq A, Bagante F, Sahara K, et al. Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery. *J Gastrointest Surg*. 2020;24(8):1843–51.
10. Nudel J, Bishara AM, de Geus SWL, Patil P, Srinivasan J, Hess DT, et al. Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database. *Surg Endosc*. 2021;35(1):182–91.
11. Shi HY, Lee KT, Lee HH, Ho WH, Sun DP, Wang JJ, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. *PLoS One*. 2012;7(4): e35781.
12. Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. *Stat Methods Med Res*. 2018;27(11):3271–85.
13. Wise ES, Amateau SK, Ikramuddin S, Leslie DB. Prediction of thirty-day morbidity and mortality after laparoscopic sleeve gastrectomy: data from an artificial neural network. *Surg Endosc*. 2020;34(8):3590–6.
14. Pera M, Gibert J, Gimeno M, Garsot E, Eizaguirre E, Miró M, et al. Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. *Ann Surg*. 2022;276:776.
15. van den Bosch T, Warps AK, de Nerée Tot Babberich MPM, Stamm C, Geerts BF, Vermeulen L, et al. Predictors of 30-day mortality among dutch patients undergoing colorectal cancer surgery 2011–2016. *JAMA Netw Open*. 2021;4(4):e217737.
16. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhimman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22(1):12.
17. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350: g7594.
18. Sarker IH. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Comput Sci*. 2021;2(5):377.
19. Stam WT, Goedknegt LK, Ingwersen EW, Schoonmade LJ, Bruns ERJ, Daams F. The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: a systematic review. *Surgery*. 2021. <https://doi.org/10.1016/j.surg.2021.10.002>.
20. Lever J, Krzywinski M, Altman N. Model selection and overfitting. *Nat Methods*. 2016;13(9):703–4.
21. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep*. 2019;49(1):15–21.
22. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
23. Li X, Yang L, Yuan Z, Lou J, Fan Y, Shi A, et al. Multi-institutional development and external validation of machine learning-based models to predict relapse risk of pancreatic ductal adenocarcinoma after radical resection. *J Transl Med*. 2021;19(1):281.
24. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
25. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*. 2019;7: e7702.
26. Mazaki J, Katsumata K, Ohno Y, Udo R, Tago T, Kasahara K, et al. A novel predictive model for anastomotic leakage in colorectal cancer using auto-artificial intelligence. *Anticancer Res*. 2021;41(11):5821–5.
27. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ*. 2020;98(4):251–6.
28. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics*. 2021. <https://doi.org/10.1136/medethics-2020-106820>.
29. Afnan MAM, Liu Y, Conitzer V, Rudin C, Mishra A, Savulescu J, Afnan M. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Hum Reprod Open*. 2021. <https://doi.org/10.1093/hropen/hoab040>.
30. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–50.
31. Lundberg CL, S-I. A unified approach to interpreting model predictions. *arXiv*. 2017;1
32. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56(5):441–7.
33. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. Cham: Springer International Publishing; 2019.
34. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer International Publishing; 2016.
35. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci*. 2019;116(32):15849–54.
36. Cao Y, Montgomery S, Ottosson J, Näslund E, Stenberg E. Deep learning neural networks to predict serious complications after bariatric surgery: analysis of scandinavian obesity surgery registry data. *JMIR Med Inform*. 2020;8(5): e15992.
37. Grass F, Storlie CB, Mathis KL, Bergquist JR, Asai S, Boughey JC, et al. Challenges of modeling outcomes for surgical infections: a word of caution. *Surg Infect (Larchmt)*. 2021;22(5):523–31.
38. Azimi K, Honaker MD, Chalil Madathil S, Khasawneh MT. Post-operative infection prediction and risk factor analysis in colorectal surgery using data mining techniques: a pilot study. *Surg Infect (Larchmt)*. 2020;21(9):784–92.
39. Gulati G, Upshaw J, Wessler BS, Brazil RJ, Nelson J, van Klaveren D, et al. Generalizability of cardiovascular disease clinical prediction models 158 independent external validations of 104 unique models. *Circ Cardiovasc Qual Outcomes*. 2022. <https://doi.org/10.1161/CIRCOUTCOMES.121.008487>.
40. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368: m441.
41. Zamanipour Najafabadi AH, Ramspek CL, Dekker FW, Heus P, Hooft L, Moons KGM, et al. TRIPOD statement: a preliminary

- pre-post analysis of reporting and methods of prediction models. *BMJ Open*. 2020;10(9): e041537.
42. Christodoulou E, Ma J, 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;110:12–22.
 43. Gravesteyn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol*. 2020;122:95–107.
 44. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7): e048008.
 45. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267–76.
 46. Kilkenny MF, Robinson KM. Data quality: “Garbage in-garbage out.” *Health Inf Manag*. 2018;47(3):103–5.
 47. Poksinska B, Jörn Dahlgaard J, Antoni M. The state of ISO 9000 certification: a study of Swedish organizations. *TQM Mag*. 2002;14:297–306.
 48. World Health O. Guide to the health facility data quality report card. 2018. p. 1
 49. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Netw Open*. 2021;4(3): e212240.
 50. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak*. 2020;20(1):280.
 51. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155(2):148–58.
 52. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Health*. 2021;3(9):e599–611.
 53. Jungmann F, Jorg T, Hahn F, Pinto Dos Santos D, Jungmann SM, Düber C, et al. Attitudes toward artificial intelligence among radiologists, it specialists, and industry. *Acad Radiol*. 2021;28(6):834–40.
 54. O’Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robotics Computer Assisted Surg*. 2019;15(1): e1968.
 55. Ho CWL, Soon D, Caals K, Kapur J. Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clin Radiol*. 2019;74(5):329–37.
 56. Mu W, Liu C, Gao F, Qi Y, Lu H, Liu Z, et al. Prediction of clinically relevant pancreatico-enteric anastomotic fistulas after pancreatoduodenectomy using deep learning of preoperative computed tomography. *Theranostics*. 2020;10(21):9779–88.
 57. Adams K, Papagrigoriadis S. Creation of an effective colorectal anastomotic leak early detection tool using an artificial neural network. *Int J Colorectal Dis*. 2014;29(4):437–43.
 58. Cave S, Nystrup R, Vold K, Weller A. Motivations and risks of machine ethics. *Proc IEEE*. 2018;107(3):562–74.
 59. de Almeida PGR, dos Santos CD, Farias JS. Artificial intelligence regulation: a framework for governance. *Ethics Inf Technol*. 2021;23(3):505–25.
 60. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. 2019;27(3):491–7.

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