Shapey *et al. Art Int Surg* 2023;3:1-13
DOI: 10.20517/ais.2022.31

Artificial Intelligence Surgery

Review

Open Access

Machine learning for prediction of postoperative complications after hepato-biliary and pancreatic surgery

Iestyn M. Shapey[1,2], Mustafa Sultan[3]

[1]Department of Pancreatic Surgery, St James's University Hospital, Leeds LS9 7TF, UK.
[2]Faculty of Medicine and Health, University of Leeds, Leeds LS9 7TF, UK.
[3]Manchester University NHS Foundation Trust, Manchester M13 9PT, UK.

Correspondence to: Dr. Iestyn M. Shapey, Department of Pancreatic Surgery, St James's University Hospital, Beckett Street, Leeds LS9 7TF, UK. E-mail: iestyn.shapey@nhs.net

How to cite this article: Shapey IM, Sultan M. Machine learning for prediction of postoperative complications after hepato-biliary and pancreatic surgery. *Art Int Surg* 2023;3:1-13. https://dx.doi.org/10.20517/ais.2022.31

Received: 30 Sep 2022 **First Decision:** 6 Dec 2022 **Revised:** 16 Dec 2022 **Accepted:** 5 Jan 2023 **Published:** 31 Jan 2023

Academic Editors: Henry A. Pitt, Takeaki Ishizawa **Copy Editor:** Ke-Cui Yang **Production Editor:** Ke-Cui Yang

Abstract

Decision making in Hepatobiliary and Pancreatic Surgery is challenging, not least because of the significant complications that may occur following surgery and the complexity of interventions to treat them. Machine Learning (ML) relates to the use of computer derived algorithms and systems to enhance knowledge in order to facilitate decision making and could be of great benefit to surgical patients. ML could be employed pre- or peri-operatively to shape treatment choices prospectively, or could be utilised in the post-hoc analysis of complications in order to inform future practice. ML could reduce errors by drawing attention to known risks of complications through supervised learning, and gain greater insights by identifying previously under-appreciated aspects of care through unsupervised learning. Accuracy, validity and integrity of data are of fundamental importance if predictive models generated by ML are to be successfully integrated into surgical practice. The choice of appropriate ML models and the interface between ML, traditional statistical methodologies and human expertise will also impact the potential to incorporate data science techniques into daily clinical practice.

Keywords: Machine Learning, artificial intelligence, hepatic surgery, pancreatic surgery

INTRODUCTION

Machine Learning (ML) relates to the use of computer-derived algorithms and systems to enhance knowledge in order to facilitate decision making. In surgery, ML has the potential to shape clinical decision making and the management of postoperative complications in three ways: (a) by using the predicted probability of postoperative complications or survival to determine and guide optimal treatment; (b) by identifying anomalous data and patterns representing high-risk physiological states during the perioperative period and taking measures to minimise the impact of the existing risks; (c) to facilitate post-hoc identification of physiological trends, phenotypic patient characteristics, morphological characteristics of diseases, and human factors that may help alert surgeons to relevant risk factors in future patients. Here we aim to review the potential clinical relevance of ML to improving the prediction of postoperative complications in hepato-biliary and pancreatic surgery.

THE CURRENT LANDSCAPE OF PREDICTING POSTOPERATIVE COMPLICATIONS

Preoperative prediction

The occurrence of postoperative complications in pancreatic surgery is a major determinant of outcomes, not least because of the impact of complications on the non-completion of adjunctive therapies. Predicting postoperative complications prior to making a commitment towards surgical therapy is important because it has the potential to change the sequence of therapies provided and the options considered. Several decision making dilemmas exist in pancreatic surgery, which include debates surrounding upfront chemotherapy *vs.* upfront surgery for pancreatic ductal adenocarcinoma (PDAC), fast-track surgery *vs.* preoperative biliary drainage for head of pancreas tumours, and parenchymal preserving *vs.* oncological resection in small neuro-endocrine tumours amongst many others[1-3]. It is also challenging to determine the resectability of malignant disease of the pancreas and ML could help play a role in reducing futile surgery - this could have a beneficial impact on patients in terms of reducing avoidable morbidity on the one hand and maximising healthcare resources on the other. There is also a need to improve the interpretation of complex multivariable patterns that represent clinical response to chemo- and immuno-therapies so that the treatment regimens and the timing of surgery could be optimised.

In liver surgery, accurate preoperative prediction of post-hepatectomy liver failure and the functioning liver remnant (FLR) could change decision making by supporting the use of adjunctive methods to increase the FLR or by counselling against higher-risk surgery in favour of other lower-risk therapeutic options (e.g., ablation or hepatic artery pump chemotherapy) where the difference in outcome may be equivocal[4-8]. ML could help identify patients better suited to more aggressive therapeutic options such as transplantation and predict which grafts and recipients are at higher risk of failure, immune rejection and mortality.

In considering the potential application of ML to predict postoperative complications prior to surgery, it is helpful to appreciate the limitations of existing models, which are primarily based on regression analyses. Three important limitations of regression based scoring systems that are commonly encountered include: (a) insufficient statistical power, often arising when the number of recorded events relative to outcomes is low; (b) when the traditional rules of frequentist classical statistics are not met, e.g., the 10-to-1 rule of 10 events for each variable included in a multivariable model; (c) where reporting of the area under the curve (AUC) is not accompanied by the standard error and p-value when making direct comparisons between models; (d) where a new variable is added to existing prediction models, but the discretional value of the additional variable is not evaluated through techniques such as Net Reclassification Improvement[9]. Regression models have struggled to translate data related to predictor variables into robust and reliable tools to improve decision making in "real-world" situations[10].

Perioperative prediction

Using perioperative data to pre-empt postoperative complications is not a new concept, and is fundamental to contemporary management of postoperative surgical patients. At an elementary level, clinicians use mental models such as recognition primed decision making, critical decision methods, and data frame theory[11-13]. These models of decision making are the framework for what is more commonly described as "expertise" or "experience". Such mental models, although often correct, are open to error, misuse or misdirection[11-13]. In the search for additional data in support of a specified hypothesis (sensemaking), individuals may be drawn along an erroneous path and misattribute data to the wrong association or cause. It is easy to fall into the cognitive trap of "explaining away" the association between poor outcomes and technical errors, or to over-interpret the significance of an adverse event in a patient whose morbidity may have little to do with the surgeon themselves. The potential value of ML, therefore, to objectively identify anomalous data and high-risk physiological patterns is of great importance. Cognitive bias may also lead surgeons to change a technical approach when no change is warranted, and vice versa.

One method of pre-emptively identifying and pro-actively addressing potential complications is the use of electronic app-based clinical algorithms, as reported by the PORSCH trial in pancreatic surgery[14]. In this randomised controlled trial of best practice after pancreatic resection in the Netherlands, algorithm-based care was used to determine when to perform an abdominal CT, radiological drainage, start antibiotic treatment, and remove abdominal drains. The algorithms described in this study represent at a human level what computers seek to achieve at a digital level. The value of algorithms of optimal perioperative care is illustrated by a significantly lower rate of the primary outcome (bleeding that required invasive intervention, new-onset organ failure, and death either during admission or within 90 days after resection) in the intervention group utilising the algorithm (adjusted RR 0.48, 95%CI: 0.38-0.61; $P < 0.0001$). It is also important to consider how ML algorithms could improve the prediction of postoperative complications above and beyond existing optimal systems and human-derived algorithms. Moreover, defining the key outcome of interest, e.g., failure to rescue rather than new-onset organ failure *per se*, is of paramount importance in shaping the way that ML will interact with clinical practice.

Modified Early Warning Scoring (MEWS) systems exist to identify and pre-empt clinical deterioration, and are based on basic physiological parameters such as heart and respiratory rate, blood pressure, oxygen saturation and requirement, and neurological status[15-16]. In many healthcare systems, MEWS systems can be set at certain thresholds to trigger pre-determined actions by clinical staff, for example, the automated review of a patient by a critical care outreach team. Such systems have been shown to have a beneficial impact on medical patient care by reducing the rate of in-hospital cardiac arrest[17-19]. The absence of individual patient context to the interpretation of MEWS data outputs (e.g., heart rate and beta-blockade or athleticism) represents a critical limitation, as does the non-identification of critical junctures that arise from reviewing isolated data outputs rather than appreciating the subtleties of data trends (e.g., swinging pyrexia). ML could help address the deficiencies in existing systems: (a) by identifying anomalous data that does not trigger an automated or human system; (b) by relating biomarker data to electronic health and prescribing records; and (c) by alerting clinicians to concerning clinical note entries through free-text associations.

Currently, the practical application of ML to perioperative care is limited by multiple stumbling blocks. These include: (a) the accuracy of alerts and the potential of spurious data to divert attention; (b) real-time delivery of alerts in a manner that could change clinical practice; and (c) convergence of data points and gate-keeping over which data ought to be considered relevant. In due course, these limitations could each be addressed by the regular auditing and quality control of ML systems, by automating real-time calculations and subsequent alerts to accompany each new piece of data, and by utilising multi-faceted and integrated electronic patient records.

Post-hoc prediction

Reviewing specified cases that experience mortality or significant morbidity is a long-standing feature of most contemporary surgical departments. However, the systematic collection of data according to pre-defined criteria and data variables is a relatively new concept that is gaining popularity. The National Surgical Quality Improvement Programme (NSQIP), championed by the American College of Surgeons, provides a structured framework from which to capture and analyse relevant data. NSQIP uses a standardised Participant Use File to collect data at the individual patient level and can be analysed according to the procedure[20]. Failure to rescue is an important binary outcome variable that is collected and reported by NSQIP and reflects the inability to identify and ameliorate postoperative complications. Meanwhile, in the UK, O'Reilly *et al.* showed that the process of instituting a prospective quality improvement programme was a significant driver behind a reduction in postoperative complications[21]. In this instance, granular data using standardised definitions of postoperative complications as agreed by the International Study Groups of Liver Surgery and Pancreatic Surgery[22-27] were prospectively collected and validated in a weekly meeting of senior HPB surgeons. Moreover, adoption of the Comprehensive Complication Index (CCI) as a continuous outcome variable representing the full and broad range of postoperative complications facilitates a standardised tool for reliable comparison amongst cohorts[28]. The success of the Dutch Pancreatic and Hepatobiliary National Audits in providing a data platform from which to perform practice changing research illustrates the potential for machine learning methods to tap into rich data repositories that could help improve outcomes[29-30].

Existing quality improvement and audit programmes highlight some important lessons that require due consideration prior to instituting ML as an integral part of the analysis of postoperative complications. First, variables and outcomes should only be reported according to clearly agreed definitions, while prospective validation of recorded data is essential in order to ensure the accuracy and integrity of ML analyses. Second, a mixture of data forms that include qualitative and quantitative outcomes (both binary and continuous) are necessary in order to capture the true impact of surgical care on patient experience. Third, measures of optimal outcomes (e.g. return to normal physiological function, and length of stay adjusted for the complexity of surgery) should be included alongside complication outcomes. Effective quality improvement mandates both the reduction of errors, deriving from the analysis of complications, and an increase in insight, deriving from the analysis of best practices. It can be challenging to gain consensus on best practice outcomes because patients, populations and health systems are very heterogenous groups. Nonetheless, it is vitally important because the minimisation of complications is associated with improvements from multiple marginal gains, whereas increasing insight can contribute to step-wise positive changes but that occur on a much less frequent basis. In the absence of detailed attention to the validity of data inputs and outcomes, the contribution of ML to quality improvement is likely to be, at best, irrelevant, and at worse, damaging to patient well-being.

Bile duct injuries occurring during minimally invasive cholecystectomy remain a problematic issue. The advent of minimally invasive surgery, including robotic systems with three-dimensional visualisation, has facilitated the opportunity for high-quality recording of surgical procedures. Artificial intelligence-assisted post-hoc review of 290 laparoscopic cholecystectomies demonstrated the ability to accurately (0.95[+/-0.06]) and specifically (0.98[+/-0.05]) identify "No-Go" zones that were representative of hazardous anatomical regions associated with a higher probability of bile duct injury. However, the technology suffered from a much lower rate of sensitivity (0.80[+/-0.21]). In this instance, the discrepancy between sensitivity and specificity is quite important, because the former has the capacity to identify a potential injury before it occurs and thereby prevent it, whereas the value of the latter lies more in confirming whether an injury may

have occurred[31]. It is reported that in due course, machine learning analysis could be incorporated in real-time.

MACHINE LEARNING, METHODOLOGY AND DATA

The frequentist approach to statistical analysis has been the most commonly used approach to understanding and interpreting data in surgical care. Its broad philosophy is to consider, within the context of narrow rules and tight assumptions, the likelihood of achieving the same result if a test were to be repeated a given number of times. Different approaches to the analysis of data have recently gained favour; for example, the Bayesian approach, which is based on the application of pre-existing data to the consideration of the *a-priori* (by theoretical deduction) conditional probability of a future event occurring. The Bayesian approach represents a far more logical and intuitive approach to statistical analysis that is highly relevant to the understanding of postoperative complications, but is currently under-utilised. In contrast to the classical approach to statistical analysis, ML takes the relative certainty of known variables and outcomes and applies algorithms to better appreciate the relationship between them. All algorithms, regardless of their classification as frequentist statistical or ML methodology, have rules and prerequisites that need to be followed. Consequently, the scientific basis for utilising a certain ML methodology ought to be outlined on each occasion, lest the validity of the work performed should be challenged.

It is helpful to distinguish between algorithms that require supervision, where clearly labelled or defined data is selected for the model, *vs.* unsupervised algorithms where the algorithm labels the data and seeks to determine the relationships between them. Reinforced learning describes a situation where the machine (i.e. a computer or robot) *automatically* processes the data for the first time and adapts its algorithms accordingly. Table 1 provides an overview of the potential application of the various ML methods, their strengths and limitations, to improve our understanding and prediction of postoperative complications. While it can be challenging to appreciate the mathematical equations that relate to the various ML algorithms, many of them are named according to everyday aspects of life that illustrate their methodology. For example, decision trees start with a trunk (i.e. the problem, or presenting state) and culminate in a series of branches that represent the various options and their associated probability of the outcome in question (e.g. survival). Random forests, therefore, represent the amalgamation of multiple trees in a given scenario. Neural networks are described in a manner that represents the neurons and synapses (i.e. nodes) in the human nervous system with the overall aim of replicating the higher functions of a human brain, albeit at a digital level.

The accuracy of ML rests on the reliability of the data entered, which comes in many forms and can be handled in many ways. In the "real world", missing data is a big problem and can be addressed, most commonly, by imputation where a value is inferred to the missing data according to the distribution of existing data. There are various methods for imputing data; modal - using the modal[most frequent] data point; multiple - by creating multiple versions of the same dataset and attributing different values from within the given distribution to the missing data, and calculating the mean value from the multiple data sets; iterative - where multiple variables are taken into consideration together in order to provide an imputed value; and arbitrary - which provides a random value from within a pre-defined range. There is also the option of removing the subjects from a data set where there is missing data, but this is infrequently advised in large and complex datasets with significant amounts of missing data. The handling of data is of critical importance because some ML algorithms cannot be legitimately performed if there is considerable missing data or if it has been addressed in a certain way. Likewise, if the outcomes have not been labelled according to clear definitions, then the validity of the results could be questioned.

Shapey *et al. Art Int Surg* 2023;3:1-13 | https://dx.doi.org/10.20517/ais.2022.31

Table 1. Machine Learning methodologies and their potential application in predicting complications following HPB surgery

Methodology	Outcome data type	Statistical assumptions	Strengths	Limitations	Optimal phase	Potential clinical application in prediction of postoperative complications in HPB surgery
Supervised models						
Linear regression	Continuous	Normal distribution of dependent variables Linear (diagonal) relationship between dependent and independent variables Observations are independent of each other Variance of residuals is the same irrespective of the value of the independent variable	Easy to execute	Poor predictive power Minimal 'tuning' of learning parameters	Postoperative	To appreciate the relationships between potential predictors and complications and also the inter-predictor relationships
Logistic regression	Binary	Linear (diagonal) relationship between dependent and independent variables Normal distribution of continuous independent variables Observations are independent of each other	Easy to execute	Poor predictive power Ability to accommodate missing, outlying or co-linearity between data Minimal 'tuning' of learning parameters	Postoperative	To appreciate the relationships between potential predictors and complications and also the inter-predictor relationships
Support vector machines	Nominal	Linear and non-linear distributions	Accommodates non-linear data Deals with outliers easily	Slow processing of very large datasets Poor performance where the distinction between there is some overlapping of outcomes	Postoperative	To identify variables associated with postoperative complications algorithms that require data from known predictors
Decision trees	Continuous Nominal (better)	Non-linear relationship (along parallel axes)	Minimal impact of missing values Easy to understand, interpret and visualise	Rely on both quality and quantity of data Small changes to data can have a big impact on the tree Variables included in the tree need to be known predictors	Preoperative	To assist treatment decision making according to the probability of a single complication or outcome
Random forest	Continuous Binary	Variables included in the analysis need to known predictors	Minimal impact of missing values or outliers Amalgamates multiple decision trees to limit errors from a single tree Easy to understand, interpret and visualise	Trees within the forest need to be discrete and not correlated	Preoperative	To consider the best treatment options by balancing the cumulative probability of individual risks and weighing up the overall benefits of treatment choices
Naive Bayes algorithm	Binary	Each variable is considered equal	Fast to execute Easy and intuitive to interpret	Reliant on accurate training data Can utilise free text data	All	Identifying triggers/red-flag features of clinical deterioration from numerical data (e.g. biomarkers, observations) and also by screening electronic health records for text "triggers/flags"
Unsupervised models						

Clustering (e.g. K-means)	Continuous Ordinal	Assumes that: - Clusters are spherical (i.e. the variance of the distribution) - All variables have similar variance - All clusters are of similar size (i.e. observations)	Easy to use and interpret Accommodates large amounts of data, including unlabelled data	Outcomes are specific to the time of analysis and data included Small changes in data will impact the outcome Reproducibility is limited Includes all data in the cohort and cannot easily adjust for outlying data	Perioperative	Anomaly detection
Principal components analysis	Continuous Nominal	Data must be standardised and scaled prior to analysis	Accommodates very large data sets with wide variations Excludes highly correlated data which does not facilitate decision making Helps understand and visualise very complex data	Prone to remove data with low variance Some data may be lost in the process of maximising	Perioperative	Identifying significant and relevant changes in biomarkers which are often highly correlated (e.g. liver enzymes/function tests, inflammatory cytokines or clinical observations)
K nearest neighbour	Nominal	None	Easy to perform Simple to understand No statistical assumptions required Responds well to new data	Requires accurate and complete data Does not easily accommodate large and complex datasets	Perioperative	Real-time identification and classification of complications according to agreed definitions (e.g. ISGPS, ISGLS)
Boosting (e.g gradient or XG boosting)	Ordinal	Data must be ordinal Assumes that datasets are incomplete (i.e. missing data) Categorical variables must be converted into numerical data	Fast execution and interpretation Minimal impact of outliers Good model performance	Difficult to interpret Challenging to 'tune' the learning parameters	Postoperative	Analysis utilising all features of an electronic health record
Supervised and unsupervised						
Neural networks	Continuous Binary	Digitalised data (i.e. not free text)	Application of established/trained models to prospective is fast and highly predictive Can easily accommodate missing data	Reliant on significant amounts of high-quality training data Training the model can be lengthy The strength of relationships between dependent and independent variables cannot be determined (unlike regression analyses)	Pre- or perioperative	Primarily to help decide optimal treatment therapies or to guide adaptations to clinical care based on a changing clinical condition (e.g. deterioration due to sepsis)

ISGLS: The International Study Group of Liver Surgery; ISGPS: the International Study Group.

CURRENT EVIDENCE OF USING MACHINE LEARNING TO PREDICT POSTOPERATIVE COMPLICATIONS

As a relatively new field of statistical analysis, there is a paucity of published evidence reporting ML-based analysis of complications following HPB surgery. Simple regression-based studies using a classical statistics approach alone have been performed for many decades and are not discussed below. Here we digest and appraise studies that have utilised more contemporary ML methodologies. Tables 2 and 3 provide a summary of the technical aspects of these ML studies.

Table 2. ML prediction of postoperative complications in pancreatic surgery

Paper	Operation	Model	Patients, centre(s)	Study	Clinical phase	Outcome	Result (aROC)
Machine learning algorithms as early diagnostic tools for pancreatic fistula following pancreaticoduodenectomy and guide drain removal: a retrospective cohort study (Shen *et al.*[53])	Pancreatoduodenectomy	CatBoost	2421, 1	Retrospective	Pre, peri & postoperative	POPF	0.81
A machine learning risk model based on preoperative computed tomography scan to predict postoperative outcomes after pancreatoduodenectomy (Capretti *et al.*[39])	Pancreatoduodenectomy	Logistic regression	100, 1	Retrospective	Preoperative	POPF	0.81
Perioperative risk assessment in pancreatic surgery using Machine Learning (Pfitzner *et al.*[54])	Pancreatectomy	Logistic regression	521, 1	Retrospective	Pre, peri & postoperative	POPF PPH, ICU readmission, death	0.37
Predicting outcomes in patients undergoing Pancreatectomy using wearable technology and Machine Learning: prospective cohort study (Cos *et al.*[45])	Pancreatectomy	Gradient boosting	48, 1	Prospective	Preoperative	Textbook surgical outcome	0.79
Risk prediction platform for pancreatic fistula after pancreatoduodenectomy using artificial intelligence (Han *et al.*[44])	Pancreatoduodenectomy	Neural network	1769, 1	Retrospective	Pre & intra-operative	POPF	0.74
Prediction of clinically relevant Pancreatico-enteric Anastomotic Fistulas after Pancreatoduodenectomy using deep learning of Preoperative Computed Tomography (Mu *et al.*[41])	Pancreatoduodenectomy	Convolutional neural network	513, 4	Retrospective (externally validated with prospective dataset)	Preoperative	POPF	0.89
The potential of machine learning to predict postoperative pancreatic fistula based on preoperative, non-contrast-enhanced CT: a proof-of-principle study (Kambakamba *et al.*[38])	Pancreatoduodenectomy	Random forest	110, 1	Retrospective cohort	Preoperative	POPF	0.95

aROC: Area under the receiving operator characteristic curve; ICU: intensive care unit; ML: Machine Learning; POPF: postoperative pancreatic fistula; PPH: post-pancreatectomy haemorrhage.

Classical statistical modelling to predict postoperative pancreatic fistula

Predicting the probability of postoperative pancreatic fistula (POPF) using classical statistical (regression) modelling has received considerable attention in the published literature[32-35]. Although these models have undergone numerous iterations and validation cycles, they continue to rely on subjective assessment of pancreatic gland texture, and intraoperative blood loss (original Fistula Risk Score), which cannot be assessed until the time of surgery. Attempts have been made to overcome these issues by using parameters determined by preoperative Computed Tomogrpahy[35-36]. Nonetheless, the reported areas under the Receiving Operator Characteristic curve (aROC) range from 0.78 in original datasets to 0.67 in subsequent cohorts aiming to validate the original studies[33,37]. The performance of the FRS is not universally consistent across patient populations from different ethnicities and cultures[37]. ML, therefore, could make a much-needed contribution to improving the reliability and reproducibility of algorithms to predict POPF.

Machine Learning modelling to predict postoperative pancreatic fistula using preoperative computed tomography

Kambakamba *et al.*'s random forest ML model showed near-perfect performance in predicting CR-POPF using preoperative CT (AUC 0.95) as compared to the FRS and a-FRS (AUC 0.80 and 0.73, respectively)[38]. Similarly, Capretti *et al.* predicted CR-POPF and postoperative length of stay using CTs from 100

Table 3. ML prediction of postoperative complications in hepatic surgery

Paper	Operation	Model	Patients, centre(s)	Study	Clinical phase	Outcome	Result (aROC)
Artificial neural network model for preoperative prediction of severe liver failure after hemihepatectomy in patients with hepatocellular carcinoma (Mai *et al.*[50])	Hemihepatectomy	Artificial neural network	353, 1	Retrospective	Preoperative	Severe PHLF	0.88
Development and validation of a Machine Learning prognostic model for hepatocellular carcinoma recurrence after surgical resection (Huang *et al.*[49])	Hepatectomy	XGBoost	7919, 2	Retrospective	Pre, peri & postoperativepostoperative	RFS	0.70
Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: a nationwide study (Shi *et al.*[46])	Hepatectomy	Artificial neural network	22,926, *multiple*	Retrospective	Pre, peri & postoperative	5-year mortality	0.89
An artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma (Qiao *et al.*[47])	Partial hepatectomy	Artificial neural network	829, 2	Prospective	Pre, peri & postoperative	Overall survival (OS)	0.83

aROC: Area under the receiving operator characteristic curve; ML: Machine Learning; OS: overall survival; PHLF: post-hepatectomy liver failure; RFS: recurrence free survival.

Italian patients using a logistic regression (LR) model, achieving AUC 0.81 and AUC 0.71, respectively[39]. These studies were limited by their retrospective nature and data from a single centre which risks *model overfitting* and limits generalizability. However, both reports showed great potential as proof-of-concept studies, and affirm recent work that has demonstrated the ability of ML to outperform human interpretation of images and recognise features inconceivable to the human eye[40].

In a larger study using the preoperative CTs of 513 patients across three centres[7], Mu *et al* developed a convolutional neural network (CNN) to predict CR-POPF that outperformed the FRS[41]. Their CNN was externally validated in a fourth centre, achieving AUC 0.89 compared to AUC 0.73 in the FRS. The CNN showed particularly higher predictive performance in the > 50% of patients deemed 'intermediate risk' by FRS (FRS 3-6). However, hepatitis B infection, which is endemic in China where the study was based, may reduce generalizability[42].

Machine Learning modelling to predict postoperative pancreatic fistula using diverse variables

ML has the potential to aggregate multiple variables and analyse complex nonlinear relationships between them[43]. This is illustrated by the report from a large retrospective Chinese study of 2421 patients undergoing pancreatoduodenectomy that utilised 59 pre-, peri- and postoperative variables in a neural network to predict POPF (aROC 0.81). A further large study of 1769 Korean patients, also undergoing pancreatoduodenectomy utislised 16 variables in a neural network model (aROC 0.74)[44]. Despite harnessing significant volumes of data, the improved performance capabilities of these models were modest compared with the performance of the FRS and aFRS. Both models incorporated variables such as intra-operative fluid status that are widely debated as to their role as predictors of POPF. It is plausible that these ML studies have uncovered variables with complex non-linear relationships that have been missed by previous classical statistical studies that assumed linearity. A number of important observations can be drawn from this data: (a) more data does not always translate into better data; (b) identification of relevant

predictor factors using classical statistical methods are reasonably robust and reliable; (c) ML might be better suited towards appreciating the complex relationships between pre-identified predictor variables and incorporating them into predictive models, rather than identification of predictor variables in the first instance; (d) ML models demonstrate greater potential as dynamic tools to guide decision making, for example, the timing of drain removal, rather than as static models that represent predicted risk at a single point in time.

Prospective machine learning prediction of complications following pancreatic surgery

Only one study prospectively studied ML prediction of post-pancreatectomy complications[45]. Cos *et al.* used a telemonitoring wearable device (*Fitbit*) to measure heart rate, step count and sleep features in 48 patients pre-pancreatectomy[45]. Combined with clinical characteristics, this activity data was used by a gradient boosting model (GBM) to predict a *textbook surgical outcome* postoperatively, outperforming the widely used ACS-NSQIP Surgical Risk Calculator (aROC: ML 0.79 *vs.* NSQIP 0.63).

Machine Learning to predict postoperative complications in hepatic surgery

In a first-of-its-kind nationwide population-based analysis of 22926 Taiwanese patients, Shi *et al.* predicted 5-year mortality post-HCC surgery using an artificial neural network (ANN)[46]. This study reported that *surgeon volume (caseload)* was the most influential factor in predicting postoperative mortality, with an AUC of 0.89. Nonetheless, the retrospective nature of this work and the absence of clinical parameters represent significant limitations that preclude the clinical utility of the model.

Machine Learning prediction of post-hepatectomy outcomes

ML approaches in hepatic surgery have mostly focused on predicting survival and recurrence post-hepatectomy in hepatocellular carcinoma (HCC). Qiao *et al.* collected prospective data on 725 patients with early HCC and predicted overall survival (OS) following minor hepatectomy using an ANN[47]. In this study, linear regression analysis was used to identify significant ($P < 0.05$) predictors, including tumor size & number, alpha-fetoprotein, microvascular invasion, and tumor encapsulation. ANN was then used to best appreciate the inter-variable relationships and develop a predictive model (aROC 0.86 - training cohort). The model was then externally validated on a separate dataset, achieving an aROC of 0.83. One limitation of ANN methodology is that the individual weightings and relationships of clinicopathological factors cannot be reported and interpreted because of the nature of the *black box algorithm* utilised by ANNs[48].

Huang *et al.* created an XGBoost model which predicted read recurrence-free survival (RFS) post-HCC resection from retrospective data collected in 7919 patients[49]. Their XGBoost model showed modest improvement over the Early Recurrence After Surgery for Liver tumour (ERASL) score in external validation (aROC; ML 0.70 *vs.* ERASL 0.67). The modest aROCs in this model highlight both the importance of high quality and prospectively validated data inputs and the impact of the chosen ML algorithm on the performance of the model. However, a unique capability reported by this study was the ability to create individualised patient risk heatmaps of tumour recurrence over time, which could inform personalised surveillance strategies.

Post-hepatectomy liver failure represents a significant postoperative complication that alters the trajectory of surgical outcomes. Mai *et al.* developed an ANN utilizing five preoperative indicators of hepatocyte function and volume (Platelet count, Prothrombin Time, Bilirubin, Aspartate Transaminase and Functional Liver Remnant) in 353 patients undergoing hepatic resection to predict severe post-hepatectomy liver failure (PHLF)[50]. This model demonstrated exceptional performance (aROC 0.88) in both training and validation cohorts and outperformed other commonly used scoring systems by considerable margins (Child-Pugh: 0.568, Model for End-stage Liver Disease: 0.608, Albumin-bilirubin: 0.627, platelet-albumin-

bilirubin: 0.584, Fibrosis index based on the 4-factor -4: 0.665, and aspartate transaminase-platelet ratio index[50].

In all three studies, generalisability of the ML models outside of a hepatitis B endemic population remains to be seen[47]. This is important, because HCC associated with hepatitis C predominates in Western populations, which also tend to be older and more obese[51].

CONCLUSION

ML shows great promise in substantially increasing the performance of statistical models to predict postoperative complications following hepatobiliary and pancreatic surgery. The accuracy, validity and integrity of data that are input into ML predictive models are central to its future success. Future studies should follow the TRIPOD-AI guidance that is currently in development[52]. ML has the potential to improve outcomes following hepato-biliary and pancreatic surgery by reducing errors through highlighting known risks of complications using supervised learning and by gaining greater insights through identifying previously under-appreciated aspects of care using unsupervised learning. The success or failure of ML to enhance clinical care will not be determined by computer science. Rather, it will be determined at a human level through our willingness to integrate the compassion of clinical care with the objectivity of data science, through our acceptance and correction of our own errors in clinical practice and data coding, and through the cultures that dominate our workplace environments and shape our attitude towards life-long learning.

DECLARATIONS

Authors' contributions

Concept and design of the paper, data collection, and authorship: Shapey IM, Sultan M

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

Both authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Versteijne E, van Dam JL, Suker M, et al. Neoadjuvant chemoradiotherapy versus upfront surgery for resectable and borderline resectable pancreatic cancer: long-term results of the dutch randomized PREOPANC trial. *J Clin Oncol* 2022;40:1220-30. DOI PubMed
2. van der Gaag NA, Rauws EA, van Eijck CH, et al. Preoperative biliary drainage for cancer of the head of the pancreas. *N Engl J Med* 2010;362:129-37. DOI PubMed
3. Hackert T, Hinz U, Fritz S, et al. Enucleation in pancreatic surgery: indications, technique, and outcome compared to standard

pancreatic resections. *Langenbecks Arch Surg* 2011;396:1197-203. DOI PubMed

4. Graaf W, van Lienden KP, van den Esschert JW, Bennink RJ, van Gulik TM. Increase in future remnant liver function after preoperative portal vein embolization. *Br J Surg* 2011;98:825-34. DOI PubMed
5. Guglielmi A, Ruzzenente A, Conci S, Valdegamberi A, Iacono C. How much remnant is enough in liver resection? *Dig Surg* 2012;29:6-17. DOI PubMed
6. Cercek A, Boerner T, Tan BR, et al. Assessment of hepatic arterial infusion of floxuridine in combination with systemic gemcitabine and oxaliplatin in patients with unresectable intrahepatic cholangiocarcinoma: a phase 2 clinical trial. *JAMA Oncol* 2020;6:60-7. DOI PubMed PMC
7. Xu Q, Kobayashi S, Ye X, Meng X. Comparison of hepatic resection and radiofrequency ablation for small hepatocellular carcinoma: a meta-analysis of 16,103 patients. *Sci Rep* 2014;4:7252. DOI PubMed PMC
8. Konstantinou I, Shapey IM, Papamichael D, de Liguori Carino N. Outcomes following potentially curative therapies for older patients with metastatic colorectal cancer. *Eur J Surg Oncol* 2021;47:591-6. DOI PubMed
9. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;30:11-21. DOI PubMed PMC
10. Bakhtiarvand N, Khashei M, Mahnam M, Hajiahmadi S. A novel reliability-based regression model to analyze and forecast the severity of COVID-19 patients. *BMC Med Inform Decis Mak* 2022;22:123. DOI PubMed PMC
11. Klein G, Phillips JK, Rall EL, Peluso DA. A data-frame theory of sensemaking. In Hoffman RR, editor. Expertise out of context. Psychology Press; 2007. pp. 118-160.
12. Klein GA. A recognition-primed decision (RPD) model of rapid decision making. In Klein GA, Orasanu J, Calderwood R, Zsambok CE, editors. Decision making in action: models and methods. Ablex Publishing; 1993. pp. 138-47. Available from: https://psycnet.apa.org/record/1993-97634-006 [Last accessed on 11 Jan 2023].
13. Klein G, Calderwood R, Macgregor D. Critical decision method for eliciting knowledge. *IEEE Trans Syst Man Cybern* 1989;19:462-72. DOI
14. Smits FJ, Henry AC, Besselink MG, et al. Algorithm-based care versus usual care for the early recognition and management of complications after pancreatic resection in the Netherlands: an open-label, nationwide, stepped-wedge cluster-randomised trial. *Lancet* 2022;399:1867-75. DOI PubMed
15. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM-INT J MED* 2001;94:521-6. DOI PubMed
16. Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N. The value of modified early warning score (MEWS) in surgical in-patients: a prospective observational study. *Ann R Coll Surg Engl* 2006;88:571-5. DOI PubMed PMC
17. Nishijima I, Oyadomari S, Maedomari S, et al. Use of a modified early warning score system to reduce the rate of in-hospital cardiac arrest. *J Intensive Care* 2016;4:12. DOI PubMed PMC
18. Beckett DJ, Inglis M, Oswald S, et al. Reducing cardiac arrests in the acute admissions unit: a quality improvement journey. *BMJ Qual Saf* 2013;22:1025-31. DOI PubMed PMC
19. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. *J Am Heart Assoc* 2018:7. DOI PubMed PMC
20. Varley PR, Geller DA, Tsung A. Factors influencing failure to rescue after pancreaticoduodenectomy: a National Surgical Quality Improvement Project Perspective. *J Surg Res* 2017;214:131-9. DOI PubMed
21. O'Reilly D, Edmiston R, Bijoor P, et al. Early experience with a hepatobiliary and pancreatic quality improvement program. *BMJ Qual Improv Rep* 2014;2:u201158.w721. DOI PubMed PMC
22. Koch M, Garden OJ, Padbury R, et al. Bile leakage after hepatobiliary and pancreatic surgery: a definition and grading of severity by the International Study Group of Liver Surgery. *Surgery* 2011;149:680-8. DOI PubMed
23. Rahbari NN, Garden OJ, Padbury R, et al. Posthepatectomy liver failure: a definition and grading by the International Study Group of Liver Surgery (ISGLS). *Surgery* 2011;149:713-24. DOI PubMed
24. Rahbari NN, Garden OJ, Padbury R, et al. Post-hepatectomy haemorrhage: a definition and grading by the International Study Group of Liver Surgery (ISGLS). *HPB* 2011;13:528-35. DOI PubMed PMC
25. Bassi C, Marchegiani G, Dervenis C, et al. The 2016 update of the International Study Group (ISGPS) definition and grading of postoperative pancreatic fistula: 11 years after. *Surgery* 2017;161:584-91. DOI PubMed
26. Wente MN, Bassi C, Dervenis C, et al. Delayed gastric emptying (DGE) after pancreatic surgery: a suggested definition by the International Study Group of Pancreatic Surgery (ISGPS). *Surgery* 2007;142:761-8. DOI PubMed
27. Wente MN, Veit JA, Bassi C, et al. Postpancreatectomy hemorrhage (PPH): an International Study Group of Pancreatic Surgery (ISGPS) definition. *Surgery* 2007;142:20-5. DOI PubMed
28. Slankamenac K, Graf R, Barkun J, Puhan MA, Clavien PA. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg* 2013;258:1-7. DOI PubMed
29. van der Werf LR, Kok NFM, Buis CI, et al. Implementation and first results of a mandatory, nationwide audit on liver surgery. *HPB* 2019;21:1400-10. DOI PubMed
30. Suurmeijer JA, Henry AC, Bonsing BA, et al. Outcome of pancreatic surgery during the first six years of a mandatory audit within the Dutch pancreatic cancer group. *Ann Surg* 2022;Online ahead of print. DOI PubMed
31. Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify

surgical anatomy during laparoscopic cholecystectomy. *Ann Surg* 2022;276:363-9. DOI PubMed PMC

32. Callery MP, Pratt WB, Kent TS, Chaikof EL, Vollmer CM Jr. A prospectively validated clinical risk score accurately predicts pancreatic fistula after pancreatoduodenectomy. *J Am Coll Surg* 2013;216:1-14. DOI PubMed
33. Mungroop TH, van Rijssen LB, van Klaveren D, et al. Alternative fistula risk score for pancreatoduodenectomy (a-FRS): design and international external validation. *Ann Surg* 2019;269:937-43. DOI
34. Roberts KJ, Sutcliffe RP, Marudanayagam R, et al. Scoring system to predict pancreatic fistula after pancreaticoduodenectomy: a UK multicenter study. *Ann Surg* 2015;261:1191-7. DOI PubMed
35. Shi Y, Gao F, Qi Y, et al. Computed tomography-adjusted fistula risk score for predicting clinically relevant postoperative pancreatic fistula after pancreatoduodenectomy: training and external validation of model upgrade. *EBioMedicine* 2020;62:103096. DOI PubMed PMC
36. Tang B, Lin Z, Ma Y, et al. A modified alternative fistula risk score (a-FRS) obtained from the computed tomography enhancement pattern of the pancreatic parenchyma predicts pancreatic fistula after pancreatoduodenectomy. *HPB* 2021;23:1759-66. DOI PubMed
37. Hayashi H, Amaya K, Fujiwara Y, et al. Comparison of three fistula risk scores after pancreatoduodenectomy: A single-institution retrospective study. *Asian J Surg* 2021;44:143-6. DOI PubMed
38. Kambakamba P, Mannil M, Herrera PE, et al. The potential of machine learning to predict postoperative pancreatic fistula based on preoperative, non-contrast-enhanced CT: a proof-of-principle study. *Surgery* 2020;167:448-54. DOI PubMed
39. Capretti G, Bonifacio C, De Palma C, et al. A machine learning risk model based on preoperative computed tomography scan to predict postoperative outcomes after pancreatoduodenectomy. *Updates Surg* 2022;74:235-43. DOI PubMed
40. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 2022;4:e406-14. DOI PubMed PMC
41. Mu W, Liu C, Gao F, et al. Prediction of clinically relevant pancreatico-enteric anastomotic fistulas after pancreatoduodenectomy using deep learning of preoperative computed tomography. *Theranostics* 2020;10:9779-88. DOI PubMed PMC
42. Chen S, Li J, Wang D, Fung H, Wong L, Zhao L. The hepatitis B epidemic in China should receive more attention. *Lancet* 2018;391:1572. DOI PubMed
43. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;19:64. DOI
44. Han IW, Cho K, Ryu Y, et al. Risk prediction platform for pancreatic fistula after pancreatoduodenectomy using artificial intelligence. *World J Gastroenterol* 2020;26:4453-64. DOI PubMed PMC
45. Cos H, Li D, Williams G, et al. Predicting outcomes in patients undergoing pancreatectomy using wearable technology and machine learning: prospective cohort study. *J Med Internet Res* 2021;23:e23595. DOI PubMed PMC
46. Shi HY, Lee KT, Wang JJ, Sun DP, Lee HH, Chiu CC. Artificial neural network model for predicting 5-year mortality after surgery for hepatocellular carcinoma: a nationwide study. *J Gastrointest Surg* 2012;16:2126-31. DOI PubMed
47. Qiao G, Li J, Huang A, Yan Z, Lau WY, Shen F. Artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma. *J Gastroenterol Hepatol* 2014;29:2014-20. DOI PubMed
48. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* 2020;172:59-60. DOI PubMed
49. Huang Y, Chen H, Zeng Y, Liu Z, Ma H, Liu J. Development and validation of a machine learning prognostic model for hepatocellular carcinoma recurrence after surgical resection. *Front Oncol* 2020;10:593741. DOI PubMed PMC
50. Mai R, Lu H, Bai T, et al. Artificial neural network model for preoperative prediction of severe liver failure after hemihepatectomy in patients with hepatocellular carcinoma. *Surgery* 2020;168:643-52. DOI
51. Shapey IM, Malik HZ, de Liguori Carino N. Data driven decision-making for older patients with hepatocellular carcinoma. *Eur J Surg Oncol* 2021;47:576-82. DOI PubMed
52. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577-9. DOI PubMed
53. Shen Z, Chen H, Wang W, et al. Machine learning algorithms as early diagnostic tools for pancreatic fistula following pancreaticoduodenectomy and guide drain removal: A retrospective cohort study. *Int J Surg* 2022;102:106638. DOI PubMed
54. Pfitzner B, Chromik J, Brabender R, et al. Perioperative risk assessment in pancreatic surgery using machine learning. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2021.pp. 2211-4. DOI