Digestive and Liver Disease 55 (2023) 704–713

Contents lists available at ScienceDirect

Digestive and Liver Disease

journal homepage: www.elsevier.com/locate/dld

Review Article

The potential role of machine learning in modelling advanced chronic liver disease

Gennaro D'Amico[a,b,1,*], Agostino Colli[c,1], Giuseppe Malizia[b], Giovanni Casazza[d,e]

[a] Gatroenterology Unit, Azienda Ospedaliera Ospedali Riuniti Villa Sofia-Cervello, Palermo, Italy
[b] Gastroenterology Unit, Clinica La Maddalena, Palermo, Italy
[c] Department of Transfusion Medicine and Haematology Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[d] Department of Clinical Sciences and Community Health - Laboratory of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", Università degli Studi di Milano, Milan, Italy
[e] Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

ARTICLE INFO

Article history:
Received 3 November 2022
Accepted 8 December 2022
Available online 30 December 2022

Keywords:
Artificial intelligence
Machine learning
Deep learning
Cirrhosis
compensated Advanced Chronic Liver Diseas (cACLD)

ABSTRACT

The use of artificial intelligence is rapidly increasing in medicine to support clinical decision making mostly through diagnostic and prediction models. Such models derive from huge databases (*big data*) including a large variety of health-related individual patient data (*input*) and the corresponding diagnosis and/or outcome (*labels*). Various types of algorithms (e.g. neural networks) based on powerful computational ability (*machine*), allow to detect the relationship between input and labels (*learning*). More complex algorithms, like recurrent neural network can learn from previous as well as actual input (*deep learning*) and are used for more complex tasks like imaging analysis and personalized (*bespoke*) medicine. The prompt availability of big data makes that artificial intelligence can provide rapid answers to questions that would require years of traditional clinical research. It may therefore be a key tool to overcome several major gaps in the model of advanced chronic liver disease, mostly transition from mild to clinically significant portal hypertension, the impact of acute decompensation and the role of further decompensation and treatment efficiency. However, several limitations of artificial intelligence should be overcome before its application in clinical practice. Assessment of the risk of bias, understandability of the *black boxes* developing the models and models' validation are the most important areas deserving clarification for artificial intelligence to be widely accepted from physicians and patients.

1. Introduction

The continuous increase in the use of Electronic Health Records (EHR) is providing large clinical databases which make it possible to quickly investigate the relations between the different types of information they contain and different types of conditions or outcomes. The study of these relations has identified a specific research area based on EHR. The interest in this type of research is steadily increasing because of the prompt availability of very large patient samples including baseline and follow-up characteristics, time to relevant clinical events and outcomes. Therefore, these studies may answer important research questions in much less time and at much less cost than prospective studies, provided that the risk of bias common in retrospective studies (mostly patient selection, attrition, blindness, and outcome assessment) may be adequately controlled.

Technological evolution in the field of computing and data storage allows the formation of huge databases (*big data*) including any type of data produced in clinical practice spanning from genome sequencing to high resolution imaging, clinical history, laboratory data, vital function monitoring, sequential clinical characteristics, treatments, time to relevant clinical events, mortality, and many others. The availability of such types of information in large patient populations is also rising the interest towards clinical research in the field of personalized clinical management, using artificial intelligence (AI) [1].

The terms Machine Learning (ML) and Deep Learning (DL) are used with reference to two different levels of complexity of AI with DL being the most complex. There is not a clearcut definition for the two terms neither a clear boundary separating them. ML may be thought about as a tool able to produce a decision rule directly from data, without any (human) predefined behavior [2,3]. As an example of human predefined behavior for a computer derived

* Corresponding author at: Viale Cavarretta 34, 90151 Palermo, Italy.
E-mail address: gedamico@libero.it (G. D'Amico).
[1] Shared first authorship

https://doi.org/10.1016/j.dld.2022.12.002
1590-8658/

prediction tool, the MELD score [4,5] was derived by a Cox regression analysis including few variables: both the type of analysis and the variables were selected by physicians in a limited patient sample. The so-called *machine learning spectrum* begins where human specification of a predictive algorithm's rules becomes so complex to be too hard to manage, although no precise limit separates "human" models from machine learning.

Deep learning is based on highly complex technologies resulting in dynamic learning algorithms and has been proposed in recent years for diagnostic models from imaging analysis and for prediction models [1,2].

A measure of the growing interest for AI in medicine is provided by the great increase in the number of publications with the terms "artificial intelligence" or "machine learning" or "deep learning" in the title, indexed in PubMed from 126 in 2010 to 11,914 in March 2022.

DL may help improving knowledge of the clinical course of diseases and its predictability resulting in valid support to decision making particularly for individualized (bespoke) medicine [6–8]. The potential applicability of DL algorithms in modeling the clinical course of cirrhosis will be summarized here by exploring areas where knowledge is still insufficient and where DL may have substantial impact.

2. Major gaps in the knowledge of the course of cirrhosis and its predictability

In its initial stage, cirrhosis is characterized by a silent course and may remain underdiagnosed for many years. This stage of the disease, termed *compensated* is associated with a median survival of ≥ 12 years [9]. When symptoms of disease progression become clinically overt, the expected median survival is reduced to 2–4 years [10,11], with a poor quality of life, need of medications and frequent hospitalizations. This is the *decompensated* stage and is defined by the occurrence of one or more of the major clinical manifestations of the disease: ascites, bleeding, hepatic encephalopathy, or jaundice. Almost all the patients with cirrhosis who die because of cirrhosis develop decompensation before death, with a small proportion dying at first decompensation while death before decompensation, for unrelated causes, is rare.

The incidence of decompensation is in the order of 5–7% [11,12] per year and depends on the severity of portal hypertension [13]. In fact, decompensation and development of gastro-esophageal varices almost exclusively occur in patients with hepatic vein portal gradient (HVPG) ≥10 mmHg, the threshold for clinically significant portal hypertension (CSPH). Among patients with CSPH the risk of decompensation is highest in those with esophago-gastric varices [13,14]. Patients with HVPG <10 and >6 mmHg have mild portal hypertension (MPH) and their risk of disease progression is the lowest, although the risk size is not yet well established.

In the last decade, non-invasive tests for fibrosis and/or portal hypertension [15], mostly the aspartate aminotransferase to platelet ratio (APRI), fibrosis-4 index (Fib-4) [16] and liver stiffness measurement (LSM), measured by liver transient elastography (TE) [17] allowed to expand the concept of liver cirrhosis to that of compensated advanced chronic liver disease (cACLD), which includes cirrhosis and chronic liver disease with advanced fibrosis at risk of decompensation even without biopsy [15]. Moreover, non-invasive tests accurately predict the risk of disease progression and mortality in cACLD [16]. LSM>15 kPa identifies patients with cACLD (sensitivity 91%, specificity 95%) [18] and LSM≥25 identifies those with CSPH (sensitivity and specificity >85%) [19]. Moreover, the combination of LSM<20 kPa and platelets count>150 × 10^9/L rules out the presence of esophagogastric varices needing treatment for the prevention of variceal bleeding with a sensitivity >95%, thus allowing a significant reduction in the indication to endoscopic screening for varices [20].

Therefore, the introduction of NITs, and particularly LSM, and of the concept of cACLD in clinical practice have substantially reduced the need for liver biopsy and/or HVPG measurement to risk stratify patients, making clinical decisions much easier and more straightforward in cACLD patients. This is particularly important because CSPH is a cornerstone in the clinical course of the disease. It is associated with a hyperdynamic circulatory syndrome, mostly in response to the splanchnic vasodilatation induced by increasing portal pressure. Parallel to the hyperdynamic circulation, bacterial translocation activates an inflammatory response resulting in progressive immune-dysfunction and reduced resistance against infections [21,22]. Increasing inflammatory activation causes further vasodilation and increase in cardiac output up to a degree where no further compensating mechanisms are possible and hemodynamic dysfunction occurs [23].

Decompensation has been considered for decades as the occurrence of one or more decompensating event independent of the modality of presentation [24,25]. In the last decade acute decompensation (AD) has been introduced as a peculiar modality of decompensation defined by the acute development of one or more major complications: first or recurrent grade 2 or 3 ascites within less than two weeks, first or recurrent acute hepatic encephalopathy in patients with previous normal consciousness, acute gastrointestinal bleeding, and any type of acute bacterial infection [26,27]. The worst expression of AD is acute on chronic liver failure (ACLF) characterized by the development of organ failures among liver, kidney, brain, circulation, coagulation, and lung. The major driver of AD is credited to be systemic inflammation with several studies showing significant increase of inflammatory markers as white blood cell count, C reactive protein or circulating levels of pro-inflammatory cytokines [28,29]. Moreover, the intensity and stability of the inflammatory activation has been reported to allow risk stratification of patients with AD [27]. However, whether inflammation is the cause or a consequence of the decompensating event, is not yet clearly defined.

There are several major issues to be clarified in the concept of AD before it is adopted to risk stratify patients in clinical practice. The first is that the reproducibility of the definition is not known. In fact, any information on AD has been drawn from patients hospitalized for decompensation with criteria for AD assessed post-hoc [26,30] and reproducibility of the criteria for hospitalization never investigated. A second major issue is that the incidence of AD is unknown, while the only available study on the incidence of ACLF reported that it was approximately 2.5/100 patient-years in outpatients with compensated cirrhosis [31]. Therefore, beyond the uncertainty on the applicability of the definition of AD, no measure is available on the impact of AD in the course of cirrhosis.

Following decompensation, disease progression may occur through further decompensation, which is associated with further reduction of expected survival and recognized by consensus [17] as a more advanced disease stage. Further decompensation is defined as (*a*) the occurrence of any of ascites, variceal hemorrhage, hepatic encephalopathy, or jaundice in a patient already decompensated, (*b*) recurrent variceal bleeding or encephalopathy or spontaneous bacterial peritonitis (SBP) or hepatorenal syndrome-acute kidney injury (HRS-AKI) or (*c*) occurrence of ascites, encephalopathy, or jaundice after recovery from bleeding but not if these events occur around the time of bleeding. In a still unpublished large multicentre study from Europe and Argentina including more than 2500 patients [17], it has been estimated that the development of further decompensation is associated with a sub hazard ratio of 1.47 (95%CI 1.26–1.72, OLT competing) in patients with a previous decompensation.

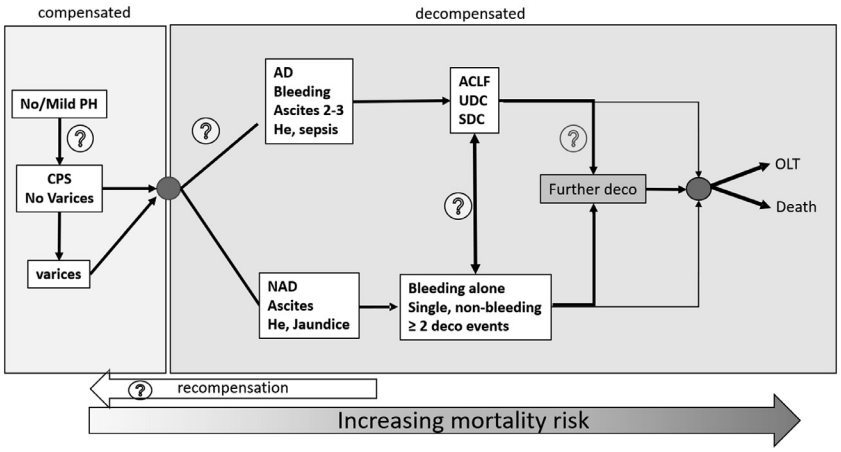

Fig. 1. Schematic representation of the model of cirrhosis. Non-Acute or Acute events cause transition from compensated to decompensated cirrhosis. Further decompensation is a next stage before the final outcomes, death, or liver transplant. The question marks represent areas where Deep Learning (DL) may improve present day knowledge. Abbreviations: pH = portal hypertension; CSPH = clinically significant portal hypertension; AD = acute decompensation; NAD = non acute decompensation; HE = hepatic encephalopathy; ACLF = acute on chronic liver failure; UDC = unstable decompensated cirrhosis; SDC = stable decompensated cirrhosis; deco= decompensating event; OLT = orthotopic liver transplant.

Etiological treatments in compensated cirrhosis may delay or prevent decompensation through the reduction of fibrosis consequent to the removal of the etiologic factor (either virus, or alcohol or obesity and metabolic alterations) [32–35], and reduction of portal hypertension up to reversion of cirrhosis [36–38], while in early decompensation it may achieve recompensation [39–42].

In summary, compensated cirrhosis progresses from an early stage of MPH to CSPH with or without esophago-gastric varices; transition to decompensation occurs through a non-acute or an acute event and first decompensation is followed by further decompensation. Death or OLT are the final outcomes. The course may be relented or reversed by etiological treatment.

A schematic representation of the course of cirrhosis is shown in Fig. 1. The major areas where we need to improve our knowledge for clinically sound decision making are the transition from MPH to CSPH, the intensity and modality of disease progression from CSPH, the impact of AD and ACLF and finally the impact of etiological treatments on recompensation.

How deep learning may substantially contribute to fill these gaps together with the relevant risk of bias will be discussed in the two next paragraphs.

3. Basic concepts of machine learning and its applicability in clinical practice

AI uses algorithms designed to adsorb information from large volume medical data and find out their relationship with a defined condition (disease or disease stage) or outcome (time to clinical events or death) to assist clinical practice. These algorithms also include self-updating instructions to improve accuracy based on regular feedback input, thereby reducing clinical error, and offering a potential for real-time diagnostic and prognostic inferences [1]. A very large variety of medical data are used for AI in medicine, encompassing screening, diagnosis, laboratory, imaging, histology, treatment, instrumental monitoring recording, data update along time, follow-up events, and outcome [43]. Special techniques [44] have also been developed to convert unstructured clinical notes (natural language processing) and intervention reports, to machine usable data.

In general, AI is a data computing approach to produce automated systems that can perform complex tasks accounting for a wide range of data combinations using several families of computing technologies. One type of these technologies, based on powerful computational ability (*machine*), is aimed to find relationships between the acquired data (*learning*) and has been termed *machine learning* (ML) [43].

Algorithms used in ML are subdivided in "*unsupervised*" and "*supervised*" learning, according to whether they are aimed at *disease characterization/diagnosis* based on patient features, or at *predicting outcome* through identification of some relationship between patient characteristics and outcome. In supervised algorithms, data are labelled according to their clinical interpretation (e.g., disease or outcome present/absent). In unsupervised models data are not labelled, and the algorithm classifies them according to common characteristics derived by specific mathematical processes [45]. The most widely used algorithms in ML are reported in Box 1.

Support Vector Machine (SVM) and Neural Networks (NN) are the most widely used supervised learning algorithms used in medical applications [1].

SVM is based on the weights to be attributed to patient characteristics to identify two groups of patients according to a relevant outcome variable and a decision boundary. The analysis algorithm is aimed at achieving the smallest classification error.

NN may be thought about as an extension of linear regression to *capture non-linear relationships* between baseline features (input variables) and an outcome of interest (label). In NN, the associations between the input variables and labels are represented through combinations of pre-specified functional arguments in a hidden layer unapparent in clinical practice because they may be hidden in the massive amount of data. In fact, the weights of the associations between the input variables and the output are adjusted at any transition across such hidden layers, aiming at minimizing the prediction error (Fig. 2, left side).

Deep learning may be considered as a modern extension of the classical neural network technique and may be envisioned as a neural network with many more layers to detect more complex non-linear relationships between patient features and the outcome of interest. It is essentially the process of training a NN to perform a given task. Recurrent neural network (RNN) for clinical tasks and Convolutional Neural networks (CNN) for imaging analysis are among the most used techniques for medical applications of DL [45]. What increases RNN precision in identifying disease patterns or in making predictions is the fact that it does not take into consideration just the actual input but also the previous input which

Box 1
Major Machine Learning development algorithms.

Model	Characteristics
Regression (linear, logistic)	Identifies relationships between input data and output (e.g., diagnosis or outcome)
Regularized regression	Incorporates techniques to avoid overfitting in regression models
Decision tree	Classification based on splitting values of input variables, similar to a clinical algorithm
Random forest	Multiple decision trees
Support Vector Machine (SVM)	Discriminates groups of patients according to patient characteristics and to a decision boundary
Nearest neighbor	Produces predictions based on similar conditions in the training sample
Neural Network (NN)	Many hidden layers used to detect more complex non-linear relationships
Recurrent Neural Network (RNN)	Accounts for previous as well as for actual input. Uses artificial neurons with self-connections to process input sequences of arbitrary length.
Convolutional Neural Networks (CNN)	Neural network with nodes designed to resemble visual Cortices. Uses hierarchical layers of pattern detectors (artificial neurons) to detect patterns in the data
Deep Neural Networks	Multiple layers between the input and output layers

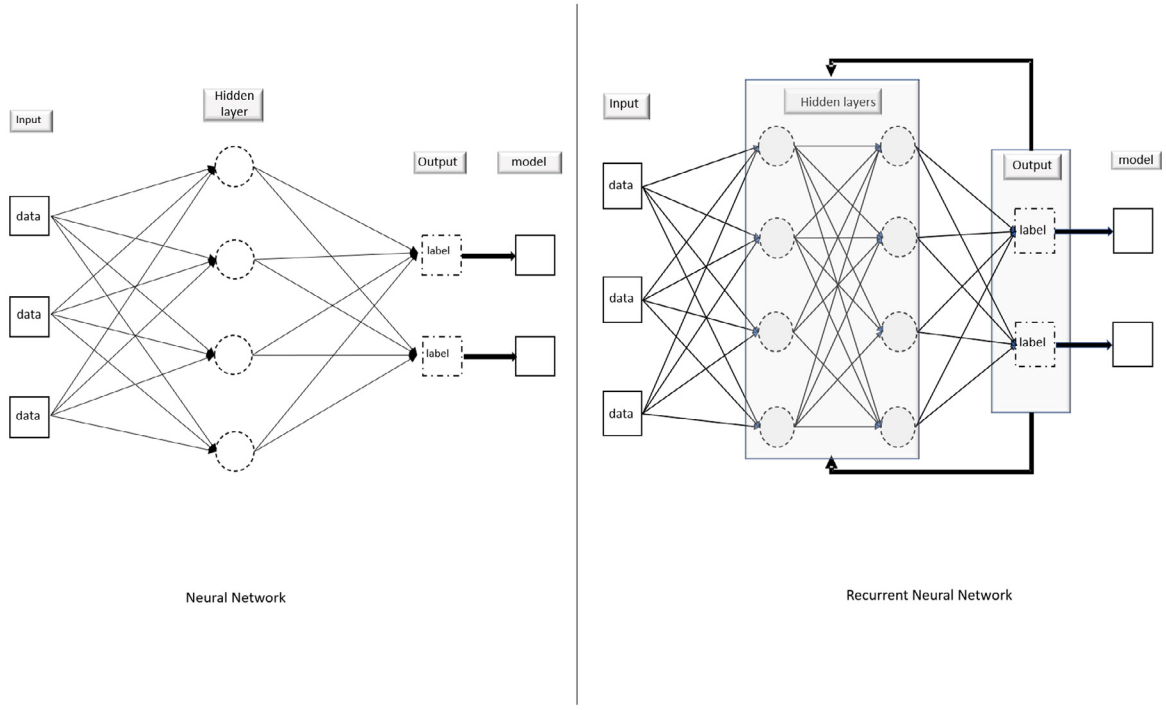

Fig. 2. Schematic representation of a neural network and recurrent neural network process through hidden layers [1].

allows it to memorize what happened previously and to adjust the covariate weights accordingly (Fig. 2, right side). Therefore, the algorithm is recursively repeated, and estimates adjusted at each new input [1].

The efficiency of DL critically depends on the training (*learning*) process. In fact, when the training dataset is not enough various or if it bears some inadvertent bias, the algorithm performance may be unsatisfactory [46]. DL algorithms are particularly suitable for complex and highly dimensional data and are mostly used in the field of diagnostic imaging, although in the last few years they are increasingly used in survival modeling.

4. Assessing the quality of machine learning prediction models studies

A careful assessment of the quality of prediction models is needed considering their proliferation and their largely inconsistent results [45,47]. For this purpose, it is suggested using PROBAST (Prediction Model Risk Of Bias ASsessment Tool), a tool designed to assess risk of bias and applicability in studies developing, validating, or updating prediction models [48].

According to the PROBAST tool, quality depends on flaws in design, conduct and analysis (i.e., risk of bias) of the study and on its applicability (i.e., the extent to which primary studies are applicable to the clinical question). The quality is assessed exploring four domains with 20 signaling questions (Table 1). Answers to each question are reported as yes/probably yes, no/probably no, no information [48]. Per each domain the ROB may be considered low if the answer to all signaling questions is "Yes" or "Probably yes". If $\geq$1 answer is "No" or "Probably no" the risk may still be judged low if specific reasons are provided supporting low risk. A high ROB should be assigned when $\geq$1 of the answers is" No" or "Probably no" in the lack of a plausible explanation minimizing the risk of bias. ROB is rated as unclear when the relevant information is missing, and the relevant argument is not judged to confer a high risk of bias.

In the first domain, patients' selection, two signaling questions aim to evaluate appropriateness of data source and in/exclusion criteria. Prospective cohorts, randomized clinical trials, and nested case–control studies are considered appropriate data sources. Case-control studies are judged at high risk of bias. Routine care registries, the data source of most ML studies, are regarded as

Table 1
PROBAST (Prediction model risk of bias assessment tool)[¶] with example[ǂ].

Domains			
1. Participants Type of potential bias Selection of participants	2. Predictors Predictors or their assessment	3. Outcome Outcome or its determination	4. Analysis Analysis
Signaling questions 1.1. Were data sources appropriate (cohort, RCT, or nested case–control study data)? [probably **yes]:** retrospective inclusion from available database	2.1. Were definitions of predictors similar for all participants? [**yes**]	3.1. Was the outcome appropriate? [**yes]**: "*we obtained all-cause mortality data from VA Vital Status File that combines information from the VA Death File, VA Compensation and Pension Benefits, Medicare, and Social Security and has a sensitivity of 98.3% and specificity of 99.8%relative to the National Death Index*"	4.1. Were there a reasonable number of outcomes observed? [**yes**]
1.2. Were inclusion/exclusion criteria appropriate? [probably **no]**: "*Our cohort included patients with cirrhosis who were seen in ambulatory clinics at 130 VA hospitals from October 1, 2011, to September 30, 2015.We included patients if they had at least 2 instances of cirrhosis …….*". Patients with at least 2 instances of diagnosis of Cirrhosis were included. We considered the exclusion of patients with only one instance of Cirrhosis as **inappropriate**	2.2. Were predictor assessments blind to outcome? [**yes**]	3.2. Was the outcome definition prespecified or standard definition used? [**yes**]	4.2. Were continuous and categorical predictors handled appropriately? [**yes**]
	2.3. Are all predictors available at the time prediction is made? [**yes**,]	3.3. Were predictors excluded from the outcome definition? [**yes**] 3.4. Were the outcome definition and assessment similar for all participants? [**yes**] 3.5. Was the outcome assessment blind to predictors information? [**yes**] 3.6. Was there an appropriate time interval between predictor assessment and outcome? [**yes**]	4.3. Were all enrolled participants included in the analysis? [probably **yes]:** "*Few laboratory values were missing in more than 5% of patients*" 4.4. Were missing data handled appropriately? [**yes**] 4.5. Was univariable analysis avoided for predictors selection? [**yes**] 4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately? [probably **yes**] 4.7. Was model assessed appropriately? [**yes**] 4.8. Were model overfitting, underfitting, and optimism in model performance accounted for? [**yes**] 4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? [**yes**]
Applicability Do included participants and setting match the question of interest? [**No**]: only patients from VA and most were older men: 96.6% males, mean age 62.7.	Do definition, assessment, and timing of predictors match the question of interest? [**no**]. "*ICD9 and drug class codes used to define predictor variables. We extracted data for serum levels of bilirubin, sodium, and creatinine and international normalized ratio performed within 1 year before and closest to the index date*" We judged the definition and timing of predictors **nor relevant nor** potentially **applicable** to the daily practice	Do outcome definition, timing, and assessment match the question of interest [**yes**]	

[¶] modified from Table 2 in ref (Wolff 2019, Ann Int Med).
[ǂ] shadowed lines report ROB assessment for reference [54] as an example of PROBAST useRCT = randomized controlled trial.

longitudinal cohorts and might be considered at high risk of bias when inclusion and exclusion criteria do not ensure inclusion of consecutively observed patients with the disease of interest.

The second domain concerns predictors. First, predictors should be defined and assessed in a similar way for all participants, especially if subjective judgment is required, such as imaging test results, as different definitions and measurements can result in biased estimate of their association with the outcome. Second, the assessment of the predictors should be made without knowledge (blinding) of the outcome. This bias is likely to occur in retrospective studies where the outcome is already known when predictors are assessed, while it may not occur in prospective studies, where the predictors are assessed before outcome verification. Third, the model should not include predictors that could not be known at the time when the model would be used. For example, to predict the risk of bleeding, using non-invasive tests in an outpatient setting, a model should not include endoscopic assessment of varices and their characteristics, as these data would be unavailable at time of outpatient visit.

The third domain explores outcome. Outcome definition may include a single (or a combination of) procedure/s or clinical judgement/s. Bias can occur when methods used to assess outcomes incorrectly classify participants with or without the outcome. First, the appropriateness of outcome assessment should be checked. Especially in routine-care registries, outcome might be assessed with suboptimal methods. Second, a prespecified or standard outcome definition should be used. Selection of an outcome definition, associated with more favorable results, may indeed overestimate the model accuracy. Third, outcomes should be assessed without information provided by predictors. If a predictor is included in the assessment of the outcome, a biased association between that predictor and the outcome is likely to be obtained (incorporation bias). Fourth, the outcome should be defined and assessed in the same manner for all study participants. The risk of bias is high when different methods are used for outcome definition (e.g., if the outcome is hepatocellular carcinoma, it may be assessed by imaging techniques, such as computer tomography or magnetic resonance or by histology of biopsy or resected specimen). The outcome should be determined without information about predictors. Knowing predictor results may influence outcome assessment and lead to biased predictive accuracy of the model, usually due to overestimation of the association between predictors and outcome. For objective outcomes (e.g. all-cause mortality) blinding is less relevant [49]. Sixth, the time interval between predictor assessment and outcome determination should be appropriate. In some cases, the time interval may be too short to capture the outcome of interest or too long, affecting the outcome definition.

The fourth domain evaluates whether appropriate analysis methods were used. First, an evaluation of the sample size and judgment of its appropriateness are required. Generally, the larger the sample size, the better. However, for prediction model studies, the number of participants with the outcome is even more important than the overall sample size. The number of participants with the outcome not only influences precision but is a potential source of bias. One of the major issues in prognostic modeling is overfitting. Usually, sample size calculations for prognostic models are based on the number of events per variable (EPV), considering the total number of predictors assessed during any stage of the prediction model process, not only those included in the final model. Prediction models developed using machine-learning techniques require a number of EPVS substantially higher (often $>$ 200) than traditional studies to minimize overfitting [50]. Second, dichotomizing continuous predictors may introduce a bias, especially when the cut-off is derived from the same dataset. In general, dichotomization should be avoided as it leads to loss of information, and an impaired predictive ability [51]. Anyway, clinically meaningful cut-off should be chosen when dichotomization is considered crucial. Third, all enrolled participants should be included in the analysis. For example, excluding participants whose predictor values were unclear (e.g., unevaluable results on imaging tests) may produce biased and overestimated accuracy [52]. In general, when the proportion of study participants excluded from the analysis is high the results are at high risk of bias. Fourth, participants with missing data should be appropriately handled, e.g., with by multiple imputation methods that are supposed to provide the least biased results. However, similarly to the patients excluded from the analysis, the risk of bias increases with a growing percentage of missing data, and the maximum acceptable percentage is hardly defined. Fifth, selection of predictors algorithms based on univariate analysis should be avoided. In many studies researchers perform a two-step analysis: a univariate analysis including all predefined predictors followed by a multivariate analysis where only the predictors with a statistically significant association (usually $P < 0.05$) at the univariate stage are included. Well-established predictors and those with clinical credibility should be included and retained in a prediction model regardless of any statistical significance. Any selection, based on statistical significance of a single predictor, might introduce bias by excluding important variables. Accordingly, a selection of potential predictors based on nonstatistical considerations should be preferred. Methods based on an a priori approach for selection of predictors are in fact considered at low risk of bias. Sixth, complexities in the data (e.g., censoring, competing risks) and underlying assumptions should properly be accounted for [53]. Seventh, both model calibration (comparing expected and actual values, i.e. model predicted probabilities and proportions of participants) and discrimination (c-index) addressing the entire range of the model-predicted probabilities should be appropriately evaluated and reported. Of note, ML algorithms require two data sets for the development stage: a training set, from which to learn parameters, and a tuning set, to adjust hyperparameters (i.e., the parameters, established before model training, that remain fixed through the training process) to avoid optimism. The model calibration and discrimination must be assessed in a third independent patients' sample (validation sample) to avoid optimistic estimates. Eighth, model overfitting and optimism in model performance should be properly considered. In ML studies, where the large sample size and EPV reduce overfitting, optimism is usually managed by adjusting hyperparameters in the tuning set. Nineth, the predictors, and their assigned weights in the final model should correspond to the results from the reported multivariable analysis.

Applicability is explored according to the domains of patients selection, predictors and outcome and depends, respectively on the matching of the included participants with the participants defined and specified by the clinical question, and on relevance and usability of predictors and outcome measurements in daily practice.

Eventually, ML prediction models, just as any other, need external validation in independent datasets obtained from different locations [45,47,48].

To illustrate the use of the PROBAST tool we show the ROB assessment of a recent study aiming to predict cirrhosis mortality [54] (Tables 1, 2). In this study we found concerns in exclusion criteria (selection domain) and in applicability (only older males were included and the definition of predictors and timing of their assessment seems not appropriate neither applicable to daily practice). Moreover, importantly, the study lacks external validation.

Another challenge is considering latent bias, that is biases waiting to happen even in fair models [55]. Adaptative models can become biased over time or according to change in the context, disproportionately benefiting individuals who already experience privilege or missing the interests of individual patients or the

Table 2
Overall assessment of the risk of bias according to the PROBAST tool for a recently published prediction model for mortality in cirrhosis derived from a ML approach [54].

Author, year	ROB				Applicability			Overall	
	participants	predictors	outcome	analysis	participants	predictors	outcome	ROB	applicability
Kanwal, 2020	-[a]	+	+	+	-[b]	-[c]	+	–	–

Footnotes
.PROBAST = Prediction model Risk Of Bias ASsessment Tool;
ROB = risk of bias.
+ indicates low ROB or low concern for applicability.
– indicates high ROB or high concern for applicability.
[a] inappropriate exclusion: participants with only one instance of "Cirrhosis" in the database was excluded.
[b] only patients from VA, mostly older men, were included (96.6% males, mean age 62.7).
[c] the definition and timing were judged nor relevant neither potentially applicable to the daily practice as some parameters were based on ICD9 codes and laboratory data were extracted within 1 year before and closest to the index date.

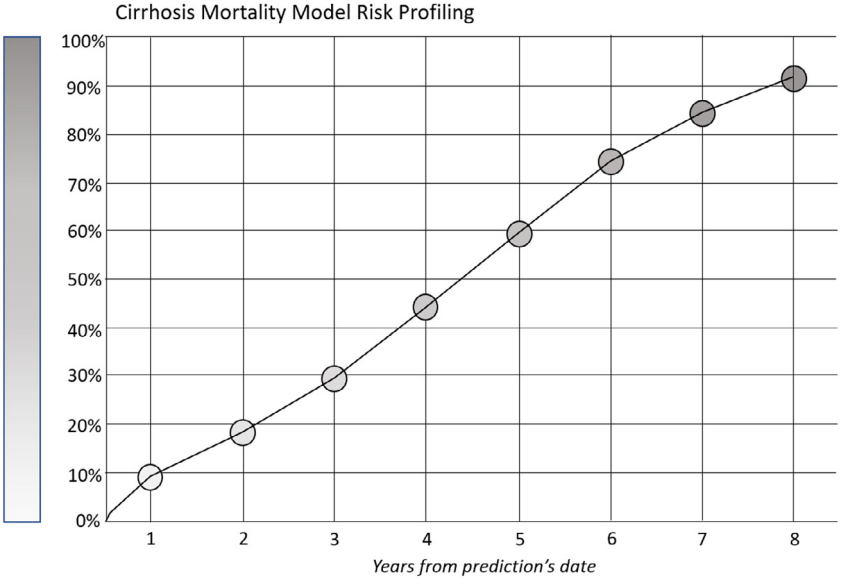

Fig. 3. Example of mortality risk diagram from 1 to 8 years for a simulated patient according to a prediction rule based on ML and refitted by a logistic analysis [ref], calculated at http://cimm.herokuapp.com/main. The graph represents the expected death risk for a cirrhotic patient 55 years old with history of chronic obstructive pulmonary disease, non-African American ethnicity, Na 133 mEq/L, bilirubin 2.4 mg/dl, 151 platelets /nL, hemoglobin 8.8 g/dL, AST/ALT ratio>2, no encephalopathy, with ascites, no HCC.

community. Addressing latent bias in AI algorithms should be seen as a patient safety issue proactively evaluated and monitored over time.

5. Potential impact of deep learning in modeling cirrhosis

DL may disclose more granularity in the course of cirrhosis through the sophisticated analyses of the huge amount of explorable data which was inaccessible until the recent past. Areas of the clinical course of cACLD where DL might provide important contribute are the disease progression and relevant risk indicators in the stage of MPH, the incidence of AD, ACLF and further decompensation. The best time for liver transplantation could be re-defined if new and more efficient prognostic tools than the MELD would be detected. The definition of decompensation per se, might be updated, being currently based on clinical signs while the role of liver function measures should be explored either as part of the definition or as predictors. Even the paradigm of decompensation as the most important risk stratifying feature might be abandoned if more efficacious indicators of disease progression are disclosed.

Several studies using DL techniques have proposed new predictive scores [56–66], although no clinical practice changing conclusions have been reached. One major obstacle in the application of DL based scores is that the complexity of covariate weighting in hidden layers, results in a sort of "black box" which may make the new models hard to understand for clinicians and patients. To overcome this "blindness" to such complex prediction tools, the covariate weights derived by DL algorithms should be converted in simpler, clinically explainable risk scores thus optimizing the trade-off between accuracy and interpretability and also making subsequent implementation easy. This approach has been explored in a very large database from the Veterans Administration, including 107,939 patients with cirrhosis [54]. The predictors identified in the most parsimonious model of the 3 ML algorithms used, were refit using maximum-likelihood discrete time-to-event logistic regression estimation. The beta coefficient per each significant parameter detected by the more complex ML model, were then used to calculate individual risks and an online risk calculator has been made available (Fig. 3).

ML may also provide more insight in treatment effect if treatment is included among candidate predictors. The analysis context recalls that of observational studies of treatment effects by propensity score matching [67]. The technique is based on the computation of *counterfactual outcomes* which consists in predicting the individual patient outcome should he or she had received the alternative treatment [68]. In this way cohorts of comparable individuals treated with two alternative treatments are generated and

treatment benefit may be computed and used in individual patient treatment decision making.

6. Machine learning limitations and challenges

A major issue with ML algorithms is that they tend to overfit data of specific training dataset, which may diminish the generalizability of the model. Control for overfitting and optimism requires a specific "tuning" step where some most important parameters (*hyperparameters*) are adjusted (tuned) in a separate "tuning set" different from the training set [45]. Moreover, the learning phase of the model (*training*) strictly depends on the type of input the model is given. If the input has not been predefined based on some strong and plausible hypothesis on biology, pathophysiology, or clinical grounds appropriate to answer specific relevant questions, the algorithm may lead to distorted conclusions [46]. Moreover, certain patient subgroups may be disregarded depending on the structure of the training set and/or data missingness.

Application of ML algorithms in clinical practice will be challenging because of several inherent unsolved issues. A first consideration is that there is not yet a satisfactory validation methodology for the predicted effects on health outcomes. Given the complexity of big data, available studies have claimed validity mostly based on a split sample technique either temporal or randomised [54]: these methods are however internal validation tools, while external validation should be performed in a similar big dataset collected in a different place. While waiting for a specific ML models validation methodology, traditional validation methods should be adopted [69] with validation samples observed in sites and including patients different than the training and tuning sets [45]. Validation may not be disregarded because while ML algorithms may be capable to uncover hidden features successfully applicable to small groups of patients, they may likewise provide spurious associations which require skilled judgement to be identified. DL models should be reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [70].

Another issue regards the vast areas of medicine uncertainty like interobserver variability, gray zone in diagnostic or outcome assessments. Such issues will be part of the information on which the learning phase of the ML process is based and will result in unintended erroneous conclusions, which may remain undetected at least for a while [46].

These limitations may be at least in part overcome adopting the "*traditional*" methods for ML based prognosis research whose challenge would be to answer "Yes" to all the PROBAST ROB signaling questions. Furthermore, the PROBAST tool, even if developed on the background of non-AI based research, can be tailored by adding additional specific signaling questions. As an example, model assessment should account for tuning sample and adjustment of hyperparameters and internal validity should be assessed in a separate participant sample, different from the training and the tuning samples. Other fields requiring specific methodology for assessing ROB might be participant selection criteria and predictors definition, that in registers need to be based on a coding system (mostly ICD9) and may have not been properly validated for the specific use relevant for a given study. Yet, any ROB assessment tool for AI, and specifically for ML based decision rules, should account for the quality of input data and consecutivity of included participants, among other specific issues.

Regarding the potential benefit of the application of AI in clinical practice it is also to be noted that no comparative studies have yet shown the effectiveness of machine learning–based decision support systems. Moreover, an important issue is represented by the implementation gap between machine learning and healthcare in spite of the level of AI performance often exceeding that of human clinicians. To overcome this problem and to allow that machine learning can add value in a real-world clinical environment, actionability (i.e. output linked to some intervention), safety (including also the evidence of efficacy in a real-world setting) and cost utility assessment have been proposed as more practical aspects to be considered and added to these sophisticated models [71].

7. Conclusions

AI is a most appealing technology for medical application, expected to provide potent support tools for clinical decision making. However, great efforts must be done to ensure transparency of the modeling process and to detect and overcome any hidden pitfall that may result in harm for the patients. Development of specific quality and performance metrics would be appropriate to ensure reliability and generalizability of DL developed clinical decision tools.

Author contributions

Gennaro D'Amico and Agostino Colli shared first authorship: review concept and design, analysis and interpretation of current relevant literature, drafting figures and tables; review supervision; drafting and revising the manuscript.

Giuseppe Malizia: review of the impact of machine learning on understanding the clinical course of cirrhosis; limitations and challenges of machine learning; revising the manuscript.

Giovanni Casazza: statistical supervision; quality assessment of machine learning based prediction models; drafting and revising the manuscript.

Conflict of interest

None declared.

References

[1] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230–43.
[2] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA - J Am Med Assoc 2018;319(13):1317–18.
[3] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347–58.
[4] Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, Ter Borg PCJ. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology 2000;31(4):864–71.
[5] Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, et al. A model to predict survival in patients with end-stage liver disease. Hepatology 2001;33(2):464–70.
[6] Mihaela V.D.S. and N. Maxfield. Survival analysis, competing risks and comorbidities [Internet]. [cited 2022 Jun 26]. Available from: https://www.vanderschaar-lab.com/survival-analysis-competing-risks-and-comorbidities/.
[7] Kneebone RL. Bespoke practice. Lancet 2017;389(10064):28–9 (London, England) [Internet]Available from. doi:10.1016/S0140-6736(16)32603-4.
[8] Kneebone RL. Making medicine bespoke. Lancet 2017;389(10064):19 [Internet]Available from. doi:10.1016/S0140-6736(16)32568-5.
[9] D'Amico G, Garcia-Tsao G, Pagliaro L. Natural history and prognostic indicators of survival in cirrhosis: a systematic review of 118 studies. J Hepatol 2006;44(1):217–31.
[10] Planas R, Montoliu S, Ballesté B, Rivera M, Miquel M, Masnou H, et al. Natural history of patients hospitalized for management of cirrhotic ascites. Clin Gastroenterol Hepatol 2006;4(11):1385–94.
[11] D'Amico G, Pasta L, Morabito A, D'Amico M, Caltagirone M, Malizia G, et al. Competing risks and prognostic stages of cirrhosis: a 25-year inception cohort study of 494 patients. Aliment Pharmacol Ther 2014;39(10):1180–93.
[12] Groszmann RJ, Garcia-Tsao G, Bosch J, Grace ND, Burroughs AKPR, Escorsell A, Garcia-Pagan JC, Patch D, Matloff DS, Gao HMRPGroup HC. Beta-blockers to prevent gastroesophageal varices in patients with cirrhosis. N Engl J Med 2005;353(21):2254–61.
[13] Villanueva C, Albillos A, Genescà J, Garcia-Pagan JC, Calleja JL, Aracil C, et al. β blockers to prevent decompensation of cirrhosis in patients with clinically significant portal hypertension (PREDESCI): a randomised, double-blind, placebo-controlled, multicentre trial. Lancet 2019;393(10181):1597–608 [Internet]Available from. doi:10.1016/S0140-6736(18)31875-0.

[14] Villanueva C, Albillos A, Genescà J, Abraldes JG, Calleja JL, Aracil C, et al. Development of hyperdynamic circulation and response to β-blockers in compensated cirrhosis with portal hypertension. Hepatology 2016;63(1):197–206.
[15] Berzigotti A, Tsochatzis E, Boursier J, Castera L, Cazzagon N, Friedrich-Rust M, et al. EASL clinical practice guidelines on non-invasive tests for evaluation of liver disease severity and prognosis –2021 update. J Hepatol 2021;75(3):659–89.
[16] Innes H, Morling JR, Buch S, Hamill V, Stickel F, Guha IN. Performance of routine risk scores for predicting cirrhosis-related morbidity in the community. J Hepatol 2022;77(2):365–76 [Internet]Available from. doi:10.1016/j.jhep.2022.02.022.
[17] de Franchis R, Bosch J, Garcia-Tsao G, Reiberger T, Ripoll C, Abraldes JG, et al. Baveno VII – renewing consensus in portal hypertension. J Hepatol 2022;76(4):959–74.
[18] Thiele M, Madsen BS, Hansen JF, Detlefsen S, Antonsen S, Krag A. Accuracy of the enhanced liver fibrosis test vs fibrotest, elastography, and indirect markers in detection of advanced fibrosis in patients with alcoholic liver disease. Gastroenterology 2018;154(5):1369–79.
[19] You MW, Kim KW, Pyo J, Huh J, Kim HJ, Lee SJPS. A meta-analysis for the diagnostic performance of transient elastography for clinically significant portal hypertension. Ultrasound Med Bio 2017;43(1):59–68.
[20] Abraldes JG, Bureau C, Stefanescu H, Augustin S, Ney M, Blasco H, et al. Noninvasive tools and risk of clinically significant portal hypertension and varices in compensated cirrhosis: the "Anticipate" study. Hepatology 2016;64(6):2173–84.
[21] Albillos A, Lario M, Álvarez-Mon M. Cirrhosis-associated immune dysfunction: distinctive features and clinical relevance. J Hepatol 2014;61(6):1385–96.
[22] Jalan R, D'Amico G, Trebicka J, Moreau R, Angeli P, Arroyo V. New clinical and pathophysiological perspectives defining the trajectory of cirrhosis. J Hepatol 2021;75:S14–26 [Internet]Available from:. doi:10.1016/j.jhep.2021.01.018.
[23] Turco L, Villanueva C, La Mura V, García-Pagán JC, Reiberger T, Genescà J, et al. Lowering portal pressure improves outcomes of patients with cirrhosis, with or without ascites: a meta-analysis. Clin Gastroenterol Hepatol 2020;18(2) [Internet]313-327.e6. Available from. doi:10.1016/j.cgh.2019.05.050.
[24] Ginés P, Quintero E, Arroyo V, Terés J, Bruguera M, Rimola A, et al. Compensated cirrhosis: natural history and prognostic factors. Hepatology 1987;7(1):122–8.
[25] D'amico G, Morabito A, Pagliaro L, Marubini E. Survival and prognostic indicators in compensated and decompensated cirrhosis. Dig Dis Sci 1986;31(5).
[26] Moreau R, Jalan R, Gines P, Pavesi M, Angeli P, Cordoba J, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology 2013;144(7).
[27] Trebicka J, Fernandez J, Papp M, Caraceni P, Laleman W, Gambino C, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. J Hepatol 2020;73(4):842–54.
[28] Arroyo V, Angeli P, Moreau R, Jalan R, Clària J, Trebicka J, et al. The systemic inflammation hypothesis: towards a new paradigm of acute decompensation and multiorgan failure in cirrhosis. J Hepatol 2021;74(3):670–85 [Internet]Available from:. doi:10.1016/j.jhep.2020.11.048.
[29] Clària J, Stauber RE, Coenraad MJ, Moreau R, Jalan R, Pavesi M, et al. Systemic inflammation in decompensated cirrhosis: characterization and role in acute-on-chronic liver failure. Hepatology 2016;64(4):1249–64 Oct 1.
[30] Trebicka J., Fernandez J., Papp M., Caraceni P., Laleman W., Gambino C., et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology: (2022) 1–9.
[31] Piano S, Tonon M, Vettore E, Stanco M, Pilutti C, Romano A, et al. Incidence, predictors and outcomes of acute-on-chronic liver failure in outpatients with cirrhosis. J Hepatol 2017;67(6):1177–84 Dec 1.
[32] Krassenburg LAP, Maan R, Ramji A, Manns MP, Cornberg M, Wedemeyer H, et al. Clinical outcomes following DAA therapy in patients with HCV-related cirrhosis depend on disease severity. J Hepatol 2021;74(5):1053–63 [Internet]Available from. doi:10.1016/j.jhep.2020.11.021.
[33] Mandorfer M, Kozbial K, Schwabl P, Freissmuth C, Schwarzer R, Stern R, et al. Sustained virologic response to interferon-free therapies ameliorates HCV-induced portal hypertension. J Hepatol 2016;65(4):692–9 [Internet]Available from. doi:10.1016/j.jhep.2016.05.027.
[34] Marcellin P, Gane E, Buti M, Afdhal N, Sievert W, Jacobson IM, et al. Regression of cirrhosis during treatment with tenofovir disoproxil fumarate for chronic hepatitis B: a 5-year open-label follow-up study. Lancet 2013;381(9865):468–75.
[35] Berzigotti A, Albillos A, Villanueva C, Genescá J, Ardevol A, Augustín S, et al. Effects of an intensive lifestyle intervention program on portal hypertension in patients with cirrhosis and obesity: the SportDiet study. Hepatology 2017;65(4):1293–305.
[36] D'Ambrosio R, Aghemo A, Rumi MG, Ronchi G, Donato MF, Paradis V, et al. A morphometric and immunohistochemical study to assess the benefit of a sustained virological response in hepatitis C virus patients with cirrhosis. Hepatology 2012;56(2):532–43.
[37] Bedossa P, Garcia-Tsao G, Jain D. Cirrhosis regression and subclassification. Surg Pathol Clin 2013;6(2):295–309 [Internet]Available from:. doi:10.1016/j.path.2013.03.006.
[38] Saffioti F, Pinzani M. Development and regression of cirrhosis. Dig Dis 2016;34(4):374–81.
[39] Pose E, Torrents A, Reverter E, Perez-Campuzano V, Campos-Varela I, Avitabile E, et al. A notable proportion of liver transplant candidates with alcohol-related cirrhosis can be delisted because of clinical improvement. J Hepatol 2021;75(2):275–83 [Internet]Available from:. doi:10.1016/j.jhep.2021.02.033.
[40] Xu X, Wang H, Zhao W, Wang Y, Wang J, Qin B. Recompensation factors for patients with decompensated cirrhosis: a multicentre retrospective case-control study. BMJ Open 2021;11(6):1–10.
[41] Cheng N, Ren Y, Zhou J, Zhang Y, Wang D, Zhang X, et al. Deep learning-based classification of hepatocellular nodular lesions on whole-slide histopathologic images. Gastroenterology 2022;162(7) [Internet]1948-1961.e7. Available from:. doi:10.1053/j.gastro.2022.02.025.
[42] Wang Q, Zhao H, Deng Y, Zheng H, Xiang H, Nan Y, et al. Validation of Baveno VII criteria for recompensation in entecavir-treated patients with hepatitis B-related decompensated cirrhosis. J Hepatol 2022:1–9.
[43] Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology. 2020;158(1) 76-94.e2.
[44] Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306(8):848–55.
[45] Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA - J Am Med Assoc 2019;322(18):1806–16.
[46] Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA - J Am Med Assoc 2017;318(6):517–18.
[47] Doshi-Velez F, Perlis RH. Evaluating machine learning articles. JAMA - J Am Med Assoc 2019;322(18):1777–9.
[48] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51–8.
[49] Savović J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association between risk-of-bias assessments and results of randomized trials in cochrane reviews: the ROBES meta-epidemiologic study. Am J Epidemiol 2018;187(5):1113–22.
[50] Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14(1):1–13.
[51] Altman DG, Royston P. The cost of dichotomising continuous variables. Br Med J 2006;332(7549):1080.
[52] Schuetz GM, Schlattmann P, Dewey M. Use of 3×2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. BMJ 2012;345(7881):1–10.
[53] Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway HADPROGRESS Group. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10(2):21001381.
[54] Kanwal F, Taylor TJ, Kramer JR, Cao Y, Smith D, Gifford AL, et al. Development, validation, and evaluation of a simple machine learning model to predict cirrhosis mortality. JAMA Netw Open 2020;3(11):e2023780.
[55] DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc 2020;27(12):2020–3.
[56] Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH. Application of artificial intelligence for the diagnosis and treatment of liver diseases. Hepatology 2021;73(6):2546–63.
[57] Su TH, Wu CH, Kao JH. Artificial intelligence in precision medicine in hepatology. J Gastroenterol Hepatol 2021;36(3):569–80.
[58] Bosch J, Chung C, Carrasco-Zevallos OM, Harrison SA, Abdelmalek MF, Shiffman ML, et al. A machine learning approach to liver histological evaluation predicts clinically significant portal hypertension in NASH cirrhosis. Hepatology 2021;74(6):3146–60.
[59] Kim HY, Lampertico P, Nam JY, Lee HC, Kim SU, Sinn DH, et al. An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B. J Hepatol 2022;76(2):311–18 [Internet]Available from:. doi:10.1016/j.jhep.2021.09.025.
[60] Ge J, Kim WR, Lai JC, Kwong AJ. "Beyond MELD" – Emerging strategies and technologies for improving mortality prediction, organ allocation and outcomes in liver transplantation. J Hepatol 2022;76(6):1318–29.
[61] Guo A, Mazumder NR, Ladner DP, Foraker RE. Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning. PLoS One 2021;16(8 August):1–12 [Internet]Available from. doi:10.1371/journal.pone.0256428.
[62] Liu Y, Ning Z, Örmeci N, An W, Yu Q, Han K, et al. Deep convolutional neural network-aided detection of portal hypertension in patients with cirrhosis. Clin Gastroenterol Hepatol 2020;18(13) 2998-3007.e5.
[63] Ioannou GN, Tang W, Beste LA, Tincopa MA, Su GL, Van T, et al. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA Netw Open 2020;3(9):e2015626.
[64] Taylor-Weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology 2021;74(1):133–47.
[65] Obeid JS, Khalifa A, Xavier B, Bou-Daher H, Rockey DC. An AI approach for identifying patients with cirrhosis. J Clin Gastroenterol 2021:34238846.
[66] Decharatanachart P, Chaiteerakij R, Tiyarattanachai T, Treeprasertsuk S. Application of artificial intelligence in chronic liver diseases: a systematic review and meta-analysis. BMC Gastroenterol 2021;21(1):1–16 [Internet]Available from. doi:10.1186/s12876-020-01585-5.

[67] Cerulli G, Ventura M. Estimation of pre- and posttreatment average treatment effects with binary time-varying treatment using Stata. Stata J 2019;19(3):551–65 [Internet]Available from. doi:10.1177/1536867X19874224.
[68] Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clin Pharmacol Ther 2021;109(1):87–100.
[69] Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68(3):279–89.
[70] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1–73.
[71] Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innov 2020;6(2):45–7.