View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Artificial Intelligence and Liver Transplant

Citation for published version:
Wingfield, L, Ceresa, C, Thorogood, S, Fleuriot, J & Knight, S 2020, 'Artificial Intelligence and Liver Transplant: Predicting Survival of Individual Grafts', *Liver Transplantation*. https://doi.org/10.1002/lt.25772

Digital Object Identifier (DOI):
10.1002/lt.25772

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Liver Transplantation

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 23. Jul. 2021

Using Artificial Intelligence for Predicting Survival of Individual Grafts in Liver Transplantation

A systematic review

Laura R. Wingfield ,[1] Carlo Ceresa,[1] Simon Thorogood,[2] Jacques Fleuriot,[2] and Simon Knight[1]

[1]Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom; and [2]The School of Informatics, Informatics Forum, University of Edinburgh, Edinburgh, United Kingdom

Abstract

The demand for liver transplantation far outstrips the supply of deceased donor organs, and so, listing and allocation decisions aim to maximize utility. Most existing methods for predicting transplant outcomes use basic methods, such as regression modelling, but newer artificial intelligence (AI) techniques have the potential to improve predictive accuracy. The aim was to perform a systematic review of studies predicting graft outcomes following deceased donor liver transplantation using AI techniques and to compare these findings to linear regression and standard predictive modeling: donor risk index (DRI), Model for End- Stage Liver Disease (MELD), and Survival Outcome Following Liver Transplantation (SOFT). After reviewing available article databases, a total of 52 articles were reviewed for inclusion. Of these articles, 9 met the inclusion criteria, which reported outcomes from 18,771 liver transplants. Artificial neural networks (ANNs) were the most commonly used methodology, being reported in 7 studies. Only 2 studies directly compared machine learning (ML) techniques to liver scoring modalities (i.e., DRI, SOFT, and balance of risk [BAR]). Both studies showed better prediction of individual organ survival with the optimal ANN model, reporting an area under the receiver operating characteristic curve (AUROC) 0.82 compared with BAR (0.62) and SOFT (0.57), and the other ANN model gave an AUC ROC of 0.84 compared with a DRI (0.68) and SOFT (0.64). AI techniques can provide high accuracy in predicting graft survival based on donors and recipient variables. When compared with the standard techniques, AI methods are dynamic and are able to be trained and validated within every population. However, the high accuracy of AI may come at a cost of losing explainability (to patients and clinicians) on how the technology works.

The number of patients awaiting liver transplantation currently outnumbers the number of donor livers available, which results in up to 20% of patients dying while on the waiting list [1,2] Because donor organs are a scarce resource, it is becoming increasingly import- ant to increase liver graft utilization and, at the same time, to ensure that the best possible outcomes can be achieved. This outcome paradigm is heavily reliant on a number of complex factors between donors, recipients, and health care providers [3,4]. Many different allocation systems have been used to find an appropriate solution to the delicate balance of factors needed to predict the best outcomes in liver organ allocation.

The current gold standard in prioritizing patients waiting for a liver in much of the world remains the Model for End-Stage Liver Disease (MELD). However, this scoring system can have conflicting results.(5,6) Other liver scoring systems include the balance of risk (BAR) score and the survival outcome following liver transplantation (SOFT), which have been validated and are currently used to assist surgeons in the decision-making process.(7) Many current scoring systems focus on the survival prognosis for the recipient, which is believed to be an extremely complex relation- ship that is nonlinear in nature.(8) However, the majority of current liver allocation models apply globally used methodologies, such as multiple regression and other linear models [9]. Therefore, achieving the most reproducible model that takes into account all donor and recipient factors would undoubtedly improve both organ use and outcomes [10,11,12].

The key to optimal organ allocation in transplantation is accurate prediction of an individual trans- plant outcome for a given set of donor and recipient variables. Such a prediction can be used as part of a matching algorithm to maximize the overall benefit from the available organ pool. It can also be used in the patient-clinician decision-making process when deciding whether to accept an organ offer.

As clinicians search for better and more accurate models predicting transplant outcomes, newer technologies are being trialed in the matching of this scarce resource. Artificial intelligence (AI), an area of computer science that encompasses intelligent and learned behaviors in computing, has been applied to many fields to produce better and more meaningful data analysis (Fig. 1) [13]. In particular, machine learning (ML), a branch of AI that extrapolates patterns and information from provided data without necessarily being explicitly instructed to do so, is quickly emerging as a vital tool in the surgical sciences for outcome prediction [14,15].

ML itself can be further categorized into different subgroups that include knowledge-based, supervised, unsupervised, and reinforcement learning methodologies [16]. For

the task of predicting outcomes following a liver transplant, supervised ML approaches are used that employ a combination of previously observed covariates (in this case, donor and recipient factors) and outcomes (in this case, observed survival times) to learn underlying relationships.

The distinction between prediction using ML and traditional statistical inference is not entirely clear-cut, but for the purposes of this review, we differentiate as follows (Fig. 1). Statistical models are those that assume an underlying probability distribution for the data generating process. A relatively small number of parameters that define the distribution are then estimated from samples of data. In the field of survival analysis, for example, Cox proportional hazards regression is a commonly used statistical model [17]. In contrast, ML approaches are more algorithmic in nature and typically have much larger numbers of associated parameters. These approaches allow complex nonlinear interactions between factors to be learned directly from data samples with few, if any, assumptions being made about the underlying distribution of the data generating process itself. Some examples of ML techniques that are commonly used for predictive tasks are artificial neural networks (ANNs), random forests, and support vector machines (SVMs) [16]. ANNs are models that use principles of statistics to build complex modeling tools using data that are nonlinear in nature, and they imitate human thinking in the way they process several data types and create patterns that are ultimately used in decision making through these neural networks [18]. Random forests create multiple decision trees that are able to sort through data and identify important variables that influence predictions or outcomes [19]. Finally, the SVM methodology organizes data by the class of variables (in a nonlinear modality), known as hyperplanes, that are able to form complex multidimensional infinite planes in space using these data [19].

To our knowledge, this is the first systematic review of AI computing techniques being used in liver transplantation to predict individual patient graft survival. The aim of this research is to provide a review of the collective evidence on AI computing techniques to predict individual patient liver graft survival when compared with the standard risk scoring tools (MELD, DRI, and SOFT).

Methods

Literature Search Strategy

Original studies on AI used to predict individual patient liver graft survival were identified by search- ing the following databases: MEDLINE, Science Direct, Springer Link, Elsevier, PubMed Central, and Cochrane databases from inception to September 11, 2019. Clinicaltrials.gov was also searched for relevant ongoing trials. No date limitations

were included within the search parameters. The following keywords were used in the search: Medical Subject Headings (MeSH) terms including “machine learning” OR “artificial intelligence” OR “neural networks (artificial)” OR “support vector machine” OR “stochastic processes” OR “Bayesian learning” OR “supervised machine learning” OR “machine learning” [title/abstract] OR “neural network” [title/abstract] OR “Bayesian learning” [title/ abstract] OR “support vector” [title/abstract] OR “machine learning” [title/abstract] OR “deep learning” [title/abstract] OR “stochastic processes” [title/abstract] AND MeSH terms “hepatic transplantation” OR “liver transplantation” OR “liver grafting” OR “hepatic transplant*” [title/abstract] OR “liver transplant*” [title/abstract] OR “liver graft*” [title/abstract] NOT editorial[publication type] or comment[publication type]. The literature search included studies published in any language. Additionally, a manual review of the reference lists of the studies obtained from the search strategy was used to identify additional relevant studies. The Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) checklist was used to include appropriate studies within this review. The study was registered on the PROSPERO database (CRD42019094865).

Selection Criteria

Studies were included if they looked at patients who received a deceased donor liver transplantation with the intervention of AI computing techniques with traditional statistical modeling to determine individual graft survival outcomes. Studies were limited to those including only liver transplantation, and those studies that involved AI and renal, cardiac, or lung transplants were excluded. Only adult transplant patients were included. Modalities of AI included within this review encompass mainly neural networks, random forests, and probabilistic graphical modeling. Overlapping study groups, although containing the same data sets, were included if they used distinctly different types of ML techniques to analyze the data or looked at different survival timescales. If studies looked at other disease areas around transplantation (i.e., hepatic carcinoma recurrence after transplantation), these studies were excluded.

No limits or exclusions were made on the number of transplant recipients or the country location of the transplants. Study design types within the inclusion criteria for this review were meta-analysis (of randomized control trials), randomized control trials, and cohort studies. Data extraction to identify ML methods, ML methodology, and associated results were recovered via a data extraction sheet. This data extraction was conducted by 2 independent researchers (LW and CC), and any discrepancies were resolved by the senior author (SK).

Quality assessment Methods

The studies within this review were assessed for quality using the Critical Appraisal Skills Programme (CASP) Checklist for cohort studies [20]. A table was constructed (Table 1) that details the 4 broad areas of quality assessment, including “validity of study,” “worth continuing work,” “results,” and “results helpful locally.” The table describes each corresponding item in the abovementioned categories with either a “yes,” “no,” or “unsure.”

Results

Literature review

Of the initial 52 studies identified via the search terms, 21 articles were excluded because they did not specifically include liver transplantation. The remaining 31 articles were selected for a further full manuscript re- view. From this group of articles, 10 examined trans- plant associated disease (no survival outcomes), 8 were excluded because upon further examination they did not include pure AI modeling, and a further 3 articles that focused solely on testing the feasibility of computing models but did not provide outcome data were excluded. For example, the article by Tusch focused on a patient decision-making pathway before and after transplant, and 1 article focused on paediatric transplant and was therefore excluded [30]. A further 10 articles were excluded because although they examined AI modalities, they looked at AI in association to transplant-related disease (i.e., recurrence of hepatocellular carcinoma in posttransplant patients), not as a tool for prediction of individualized graft survival outcomes. One article examined pediatric liver transplant survival, and therefore, it was excluded. A final 9 articles [21,22,23,24,25,26,27,28,29] were selected for inclusion within this study. Figure 2 shows the PRISMA flow diagram for paper selection.

The 9 articles relevant to the topic of AI and liver transplant outcomes included a total of 18,771 study participants (Table 2). Within these articles, the majority examined graft survival at 3 months and/or graft survival at 1 year with 8 articles examining these endpoints each. One article examined survival between 2 and 5 years. At the other end of the spectrum, some articles looked at graft survival in the very short term with 2 articles reviewing survival predictions at 30 days or fewer.

Participants

There was a mean of 2086 (range, 180-12,239) participants from all studies included in the review, and there was a tendency to analyze larger cohorts of data. For example,

Hoot et al. reviewed a cohort of 12,239 patients from American United Network for Organ Sharing (UNOS) registry data [26]. Additionally, groups were analyzed from a number of countries and transplant data sources including Spain, UNOS data from the United States, several UK centers, an Australian Liver Database, and China (Table 2). Notably, 5 of the study groups used the same data set in their research, which comprised Spanish transplantation centers with an externally validated data set from King's College London [21,22,23,24,28].

Quality Assessment of studies included in the review

The articles included were good quality overall (Table 1), with all articles reviewed including a clear research aim. A total of 77% of the articles reviewed contained results that could be used in clinical practice (n = 7). Despite these strong quality points, there were other areas that the articles could have improved upon. For instance, in almost all of the articles reviewed (n = 8), it was difficult to determine whether the results could be applied to local populations. Further areas of improvement were centered around bias reporting. Although the authors within the studies may have specifically measured outcomes to minimize bias, this was not explicitly mentioned in 3 articles [21,23,30].

Artificial Intelligence Approaches and Input Features

There is currently no gold standard in AI modeling for clinical outcomes, and a number of different approaches were trialed within our study groups. Approaches used within the articles reviewed for this article include the following: ANNs, SVM, random forests, gradient-boosted trees (GBT), and Bayesian networks. More than half of the included studies (n = 5) trialed more than 1 type of ML approach to predict the same outcome (i.e., graft survival) [21,22,25,26,30]. Despite the survival analysis setting and the existence of right-censored data, most studies approached the task as a binary classification problem at a specific time (e.g., graft survival at 1 year after transplant). As such, observations censored prior to the time of interest were excluded from the modeling process. A number of articles approached the task as one of ordinal classification (ie, the simultaneous predictions of 2 or more intrinsically ordered classes) where the classes represented failure at successive times of interest. The most commonly used type of classifier was ANNs, with 7 articles adopting this approach (Table 3) [20,21,22,25,26,27,30].

It is well established that the process of organ allocation is extremely challenging and can be influenced by many potentially interacting factors. In the 9 articles included, there was a range of donor and recipient characteristics that were examined. The number of donor, recipient, and surgery-related variables ranged from 10 to 276. Lau et al. measured the largest number of overall variables: 173 recipient variables and 103

donor recipient variables initially. They reviewed these variables and selected a top 15 inputs following feature selection [27].

Comparisons between AI modeling and standard organ risk stratification systems were made within almost all of the articles. Standard liver risk stratification scores used for comparison included MELD, SOFT, Predicting Survival Following Liver Transplantation (P-SOFT), BAR, and DRI, with MELD being the most commonly used (n = 8). Where comparisons were quantitative, a number of performance metrics were used. The most commonly used was the area under the receiver operating characteristic curve (AUROC) [20,21,24,25,30] with a C index [27] with accuracy, sensitivity, and the geometric mean of sensitivities (GMS) [26] also being reported in some cases. When reviewing the included articles, it is important to con- textualize the results in reference to the area under the curve (AUC), or C statistic, which ranged from 0.5 (showing no discrimination) to a perfect model showing a maximum value of 1. Therefore, an AUC score of 0.5 would be equivalent to a coin flip of chance. In a clinical context, models with an AUC score of 0.7 are considered a good fit, and an AUC score of 0.9 is considered an almost perfect model [31]. Ultimately, clinical judgment was used in combination with standard organ risk stratification systems including MELD scoring and previous UNOS allocation systems.

Validation of AI models used

Within the articles selected for this review, there were a variety of approaches undertaken to validate the data. However, all 9 studies validated their data sets. The most common methodology was cross-validation (n = 5) with researchers using either a 3-fold or 5-fold stratified cross-validation or a 10-fold stratified cross-validation (Table 4) [21,22,24,26,28]. Another approach taken was a training and test data set from Zhang et al., which included an 80%/20% train-test split with an additional 20% validation set created from the training set [29]. Cruz-Ramírez et al. included a 75%/25% train-test split, with multiple bootstrap samples created from each set [23,27,30]. Lau et al. used 1000 bootstrap samples with out-of-bag samples for validation and was the only group to take this approach, which is most likely due to their small sample size [27]. Interestingly Haydon et al. did not explicitly mention within their work the parameters used in evaluating the efficacy of their model, only mentioning that a separate database of 2622 patients was used for validation. They were the only group not to mention these [25].

AI compared with other predictive methods

Five of the studies directly compared the performance of ML models to some form of linear regression modelling (Table 3) 19,22,24-26,30]. Almost half of the studies within

this review did not mention specific regression methodology used (n = 4) [21,23,26,27]. Of the studies that directly compared ML models with standard regression modeling (n = 5), only 2 out of this total reported higher accuracy in ML. For instance, Briceño et al. reported their best ML model as ANNs with an accuracy of 0.91 and an AUROC of 0.82 compared with their most successful logistic regression model with an accuracy of 0.89 [21]. Although statistically speaking this is an improvement, the difference of 0.02 between traditional logistic and ML models cannot be considered a clinically significant improvement. Cruz-Ramírez et al. also compared ML techniques to logistic regression and showed similar accuracy between the 2 modalities [23]. However, they also showed an overall improved AUROC with ML. The logistic model tree ML technique had an accuracy of 0.88, whereas the best logistic regression model was 0.88. The ANN model showed a minimum sensitivity of 0.50 compared with the logistic regression of 0.03 and an AUROC of 0.57 compared with the logistic regression AUROC of 0.51 [21].

Finally, only 2 studies directly compared ML techniques to liver scoring modalities (i.e., DRI, SOFT, and BAR). Briceño et al. showed an ML AUROC of 0.82 compared with a BAR of 0.62 and a SOFT of 0.57.(21) The 0.20 difference in AUC between traditional and ML models can be considered clinically significant. Following these results, Briceño et al. [21]. suggested that their ML scoring system using ANN could be used to predict 3-month outcomes in conjunction with clinical judgment [21]. They also stated that they believed that ANNs may be the best method to combine the myriad variables involved in transplantation (i.e., donor, recipient, and others) to obtain optimal survival. They do not directly state that this methodology should replace current liver indices or linear regression models. However, they do explain that many of the current scoring modalities (DRI, MELD, SOFT, and BAR) use logistic regression analysis, which assumes a linearity among the liver transplant variables and survival. The authors point out that in reality, liver transplantation follows a nonlinear pattern, and therefore, this approach is too simplistic.

The other research to directly compare ML modalities to traditional liver scoring was from Lau et al. The ANN model from Lau et al. showed an AUROC of 0.84 compared with a DRI of 0.68 and SOFT of 0.64 [27]. These results show a difference of 0.16 in AUC values between the best ML model and the DRI, which can be considered high enough to practically warrant clinical use of one model over another. Despite the results, Lau et al. note that this initial research was a proof of concept that could potentially be used to support clinical decision making in liver transplant organ allocation. They did not outwardly suggest that this methodology should be used instead of current liver indices or linear regression modeling. However, Lau et al. suggested that their ML algorithm could be used as a tool to improve clinician confidence in using marginal organs [27].

Discussion

To our knowledge, this systematic review is the first of its kind that reviews ML methodology to predict individual graft outcome following liver transplantation. Other systematic reviews have examined ML and renal graft outcomes as reported by Senanayake et al. [33] Nursetyo et al. [34] and Sousa et al. [32] reviewed ML applications in heart, heart-lung, and kidney transplanted organs published between the years of 2009 and 2010. These limited systematic reviews in AI/ML and organ transplantation highlight a gap in the literature [33,34]. As well as limited reviews on the topic, the results have highlighted the heterogeneity in the AI techniques used within the study results. Unfortunately, only a few articles directly compared ML modeling with logistic regression and/or liver scoring systems. It would be especially useful to examine this parameter because re- searchers would be using the same data sets to compare ML performance with what is currently being used in liver scoring systems clinically.

Studies within this review by Briceño et al., Cruz-Ramírez et al., and Lau et al. demonstrated that ML modeling provided more accurate results when compared with standard regression and liver scoring modality [21,23,27]. Recently, in the United Kingdom, the liver allocation system has changed, whereby a “transplant benefit score” is generated. This system aims to provide a more in-depth score than the previous United Kingdom Model for End-Stage Liver Disease classification and comprises 7 donor and 21 recipient characteristics and is calculated using linear regression modeling measuring the difference between the AUC for the waiting list survival curve and the AUC under post-transplantation 5-year survival [35]. In time, it will be important to establish whether this new approach results in an increase in the number of life-years gained from transplanted livers and a decrease in the number of wait-list deaths. Further studies are certainly needed to select a universal AI methodology, and support from national bodies needs to be garnered. This review study highlights the growing evidence to support AI technology as a predictive tool that can be used to form an organ allocation system when com- pared with standard methods of allocation currently in use. Although AI technology is being used in other fields of medicine, it has yet to have widespread, nationally implemented programs at the time of this publication [36].

Advantages to ML Systems

In addition to the need for a nationally adopted standard AI system, the advantages of AI need to be further stressed with education of governmental and health organizations taking place. For instance, one advantage of AI methodology, especially neural networks, when compared with standard techniques is that they are dynamic and able to be trained and validated within every population. Furthermore, the more variables that are examined in terms of donor characteristics, the more precise a neural network can

be, provided enough data are available. A more accurate neural network allows for better organ allocation decisions that take into account a large number of variables, which standard programs currently may not include, or that may need a clinician's detailed review of donor and recipient parameters. To achieve this precision, however, there is a need for a standardized curation of liver transplantation data to allow widely usable, standardized models to be built to be used across different counties. Ultimately, for the individual patient, this information will help clinicians to make more informative, personalized decisions around organ acceptance and to adapt the informed consent process to the organ on offer at any one time. Finally, there are significant financial costs and regulatory constraints related to liver transplant, and as these constraints increase, it is important to have a quantitative tool to help transplant clinicians make these critical organ allocation decisions.

Challenges with Implementing ML Systems

Despite potential benefits to using ML in liver transplantation, there are also potential limitations to using this emerging technology. Ultimately, researchers using ML-based algorithms aim to present the most accurate prediction output as the data will allow. In some cases, the algorithms may include variables that, based on clinical experience and previous research, are not biologically plausible. This creates a conundrum for researchers, clinicians, and the patients, who will ultimately receive the liver transplant: if a variable is known to have no survival benefit in clinical practice, but it shows a survival benefit anyway, should it be included? This leads to the ethical issue around explainability in AI, which is described in detail later. Another limitation is that these algorithms may not have global applicability, and instead, they are often best suited to predicting outcomes based on data sets from which they were originally derived. Further challenges with ML algorithms may stem from shifting patient populations. Because algorithms are designed based on relatively static data sets in a particular point in time, there may be a new distribution in the data compared with the original data set used to train the ML algorithm, and mechanisms need to be in place to update models over time. Finally, there are logistical challenges around translating ML algorithms directly into a clinical set- ting. For instance, different health care computer systems may not easily host the programs required to run the algorithms. Final hurdles to implementation may center around clinician acceptability and implementation of algorithms when there can be considerable opacity around the algorithms themselves.

Study Strengths and Limitations

The strengths of this review include that it is the first study to systematically review all of the available literature of AI/ML techniques in liver transplant. Thus far, there has been extremely limited work in this field, and this review aims to amalgamate the current work

in this area to date. This review also discusses a timely subject, namely, ML in health care, and as this methodology is being rapidly integrated into health care systems, it is vital that the research in this area is disseminated for the transplant community. Finally, this review covers a wide breadth and depth of study participants including transplant recipients from Australia, Spain, United Kingdom, United States, and China. A total of 18,771 liver transplantation patients comprised the data sets used by researchers also making this article the largest, systematic review of its kind.

There are some limitations to this study. Many of the articles included within this review were observational studies, i.e., data being retrieved from databases. They were retrospective in nature, and to truly test the predictive power of AI, it would be ideal to have further large, externally validated prospective cohort studies. These studies could review prospective outcomes, such as 30-day graft survival as well as longer-term graft survival rates (i.e., 1 year or more). An additional limitation to this study is the high rate of heterogeneity in AI techniques used among the articles included within this review as previously mentioned. The size, heterogeneity, and quality of data sets among the studies included make direct comparisons between the studies challenging, including evaluations between regions/countries, early verses late prediction abilities of the ML algorithms, and prediction comparisons by MELD strata groups. Finally, none of the studies took a time-to-event (i.e., graft failure) approach. In clinical practice, it may be useful to have such a quantitative measure of the likelihood survival gain by accepting/ rejecting a specific organ, which would further help clinicians and patients make more informed decisions on grafts.

Explainability in ML

ML methodologies are being used to analyze data in completely new ways. This evolution in science is creating a great potential to develop clinical decision support tools that can help doctors and patients make critical decisions about health. However, the use of ML in critical health care decisions brings up several challenges. The high accuracy of ML may come at a cost of losing explainability (to patients and clinicians) on how the technology works. Some ML-based algorithms work in ways that are unknown to the creator and, therefore, cannot be explained to patients or doctors using them (which is known as a black box issue) [18]. This raises questions about accountability for such algorithms in the event that an incorrect result (i.e., incorrect liver graft survival prediction) is made. Furthermore, incomplete data sets used to train ML algorithms may cause potential biased outcomes. Finally, the use of ML poses questions around the acceptability of patients and their carer when decisions are delegated (partially) away from humans and more so to computer algorithms. On the basis of these challenges around ML algorithms, it is essential to have research into how both clinicians and patients would interpret ML-generated algorithms.

Despite challenges around the black box issue, if AI techniques were widely accepted within liver trans- plant, this could be an invaluable tool in providing clinicians with critical decision making. It would allow surgeons to make more evidence-based decision making as well as provide patients with a tool to under- stand the risk/benefit ratio of accepting a specific liver transplant. Finally, an AI model could make the liver donation system more efficient and would cause fewer organs to be discarded.

AI	**What Is the Difference Between AI and Traditional Statistical Modeling?**		**How Are Results Reported in AI? Reported Metrics**		
	AI	***Statistical Modeling***	***Accuracy***	***Minimum Sensitivity***	***AUROC***
AI is a branch of computer science that emphasizes the creation of intelligent or human-like behaviors in computer systems.	Models are algorithmic in nature. No assumptions about underlying distributions are required in many cases.	Modeling assumes an underlying distribution for the data generating process.	The ratio is given of the number of correct predictions made by the ML algorithm on a data set as compared with the total number of input data.	Also known as the true positive rate or recall, this is a numerical representation of how often the ML prediction is correct when an actual value is positive.	AUROC is one of the most common measures in ML.
Supervised ML is a branch of AI that uses algorithms that learn from existing data to uncover complex patterns and relationships that can be missed by traditional statistical models.	Potentially large numbers of model parameters must be learned from the data.	Typically, a relatively small number of model paramaters are fitted to data samples.	It works well in data sets with a relatively equal number in each classification group.	Sensitivity is calculated by true positives/(true positives + false negatives).	It is used to measure whether an ML algorithm can discriminate whether something is present or not (eg, organ survival/organ failure).
Supervised ML is increasingly used in clinical applications including diagnosis and outcome predictions.	AI is able to learn complex nonlinear relationships between factors.	Typically, modeling is unable to account for nonlinear interactions between factors.	It can be problematic to use AI in instances of rare occurrences (eg, rare disease) as the cost of misclassification of the smaller class can have high ramifications.	A perfect sensitivity score of 1.0 would indicate all relevant data were retrieved. For every clinical question, there is a benchmark (minimum) that is considered acceptable.	An AUROC of 0.5 is equivalent to a test that has no discriminatory abilities (ie, similar to chance and flipping a coin).

Fig 1. Key definitions of AI terminology

		Validity of Results		Worth Continuing work				Results		Results help locally		
		Clearly focused question?	Cohort recruited in acceptable way?	Exposure measured to minimise bias	Outcome accurately measured to minimise bias	Identified confounding factors	Follow up of subjects complete	Are results precise?	Believable results	Can results be applied to local population	Results fit with other evidence	Results may be used in practice?
Study Author	**Year**											
Briceno J.	2014	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Cruz-Ramírez M	2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Cruz-Ramírez M	2012	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Dorado-Moreno M	2017	Yes	Yes	Unsure	Yes	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Haydon GH	2005	Yes	Yes	Unsure	Unsure	Yes	Yes	Yes	Yes	No - very small population, only in Birmingham	Yes	Unsure
Hoot N.	2005	Yes	Yes	No	Unsure	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lau L	2017	Yes	Yes	Unsure	No	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Perez-Ortiz M	2017	Yes	Unsure - not mentioned	Yes	Yes	Yes	Yes	Yes	Yes	Unsure	Yes	Yes
Zhang M	2012	Yes	Yes	Unsure	Unsure	Unsure	Yes	Yes	Unsure	No	Yes	Unsure

Table 1: Modified Critical Appraisal Skills Programme (CASP) Checklist for cohort studies

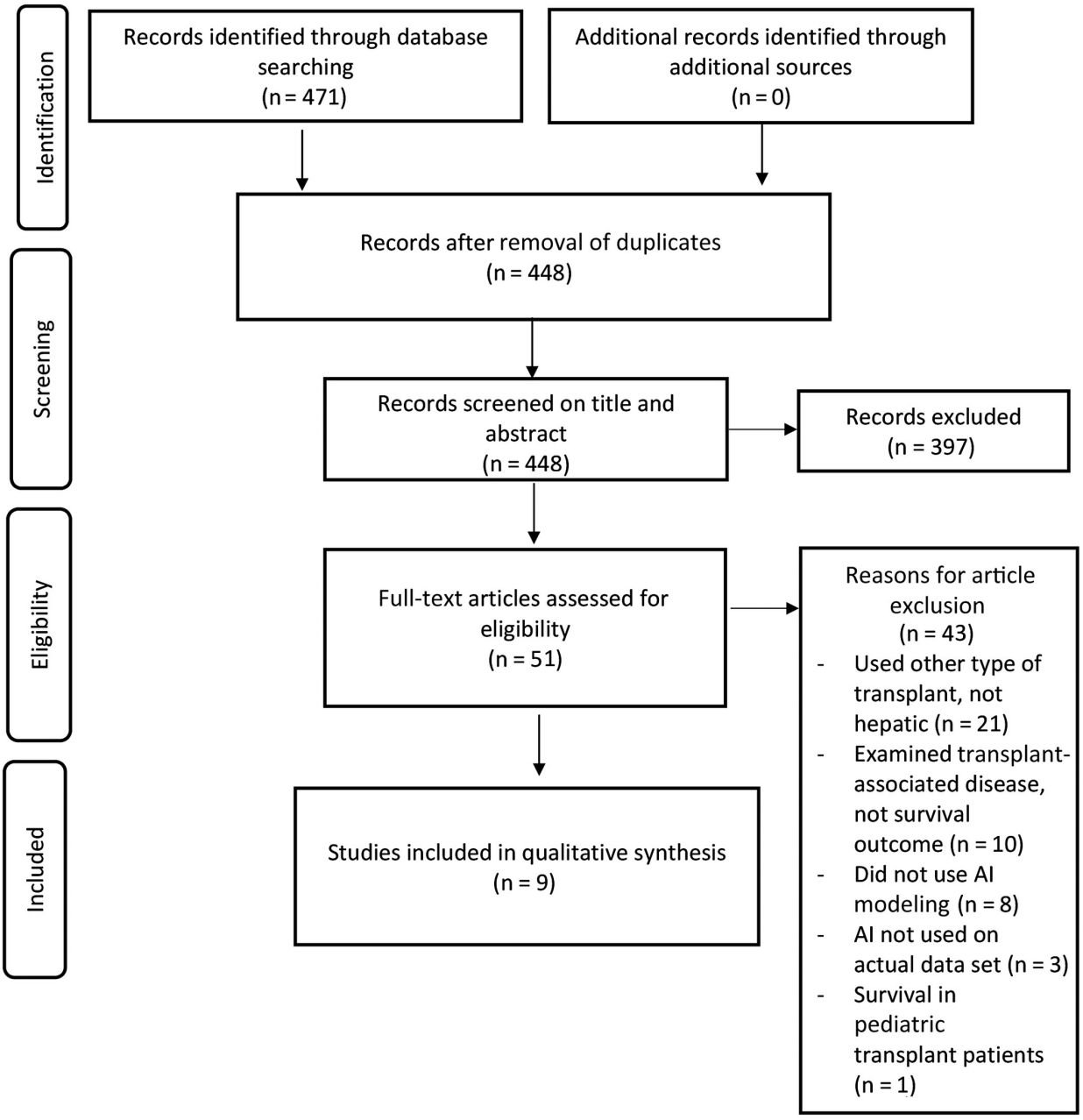

Fig 2. PRISMA flow diagram of systematic identification, screening, eligibility, and inclusion criteria.

Author	Year	Specific patient demographic group	Patient number	Average Patient Age (median)	Sex (male, number and percent)	BMI (mean +/- standard deviation)	MELD (mean +/- standard deviation)	Reason for Transplant: Alcoholic cirrhotic liver (number and percent)	Reason for Transplant: Hepatitis C Virus (number and percent)
Briceno J.	2014	11 Spanish centres	1,003	52.98 ± 9.86 (Mean +/- SD) 18-73 (range)	724 (72.2%)	26.76 ± 4.44 Range: 10.0-60.3	MELD (inclusion) 16.51 ± 6.567 (range: 1-48) MELD (at transplant) 17.35 ± 7.01 (range: 2-57)	298 (29.7%)	329 (32.8%)
Cruz-Ramírez M.	2013	11 Spanish centres	1,003	N/A	N/A	N/A	N/A	N/A	N/A
Cruz-Ramírez M.	2012	11 Spanish centres	1,001	N/A	N/A	N/A	N/A	N/A	N/A
Dorado-Moreno M.	2017	7 Spanish centres and King's College Hospital, UK	1,406	N/A	N/A	N/A	N/A	N/A	N/A
Haydon G.H.	2005	1 centre (The Queen Elizabeth Hospital Birmingham, UK)	827	median age was 52 years (range 16.5–73.5 years	479 (58%)	N/A	N/A	N/A	N/A
Hoot N.	2005	American UNOS data set from 2000-2002	12,239	Years 2000-2001; 50.8 ± 10.0 Years 2002: 51.1 ± 9.7	Years 2000-2001: 65.3% Year 2002: 68.8%	N/A	N/A	N/A	N/A
Lau L.	2017	Liver Transplant Database from Austin Health, Melbourne, Australia, from January 1988 to October 2013.	180	45.8 ± 16.8 (14-78)	52.8%	26.3 ± 4.5 (Range: 17.6-40.4)	18.2 ± 7.5 (range: 6-43)	8.9%	22.8%

Perez-Ortiz M.	2017	1 centre (King's College Hospital, UK)	822	N/A	N/A	N/A	N/A	N/A	N/A
Zhang M.	2012	1 centre (West China Hospital). Hepatic cancer patients only.	290	46.1 ± 10.8	169 (58.3%)	21.1 ± 4.4	<14: n=109 (37.6%) 14–23: n=95 (32.8%) ≥24: n=86 (29.7%)	27 (9.3%)	25 (8.6%)

Table 2: Study demographics - Recipient

Author	**Year**	**Description of ML models used**	**Supervised or Unsupervised Learning**	**Description of Regression methods used (e.g. Logistic or Cox PH)**	**Best Performing Regression method: Reported Metrics (e.g. Accuracy, Min. Sensitivity, AUC ROC)**	**Best Performing Liver Risk Index Score: Reported Metrics (e.g. Accuracy, Min. Sensitivity, AUC ROC)**	**Best Performing ML type: Reported Metrics (e.g. Accuracy, Min. Sensitivity, AUC ROC)**
Briceno J.	2014	2 x ANN Models fitted using evolutionary algorithms (NNEP)	Supervised Learning	Logistic Regression variants	Best Log. Reg.: Accuracy: 0.89	AUC ROC - BAR: 0.61 - SOFT: 0.57	Best ANN - Accuracy: 0.91 - AUC ROC: 0.82
Cruz-Ramírez M	2013	Multiple RBF-based ANNs fitted using evolutionary algorithms (MPENGSA2) Decision Tree Logistic Model Tree (LMT) SVM	Supervised Learning	Logistic Regression variants	Best Log. Reg. - Accuracy: 0.88 - Minimum Sensitivity: 0.03 - AUC ROC: 0.51	Not mentioned	Best ML - Accuracy (LMT): 0.88 - Minimum Sensitivity (ANN): 0.50 - AUC ROC (ANN): 0.57
Cruz-Ramírez M	2012	Multiple RBF-based ANNs fitted using evolutionary algorithms (MPENGSA2)	Supervised Learning	Not mentioned	Not mentioned	Not mentioned	Best ANN - Accuracy: 0.84 - Minimum Sensitivity: 0.52 - AUC ROC: 0.57
Dorado-Moreno M	2017	Cost-sensitive ordinal ANN fitted using evolutionary algorithms. Random forests, Gradient-boosted trees (GBTs), SVMs, Extreme learning machine for ordinal regression (ELMOR), Kernel discriminant leaning for ordinal regression (KDLOR)	Supervised Learning	Proportional Odds Model (POM)	GMS: 0.00	Not mentioned	Best ML - Accuracy (ELMOR): 0.85 - Geometric mean of sensitivities (GMS) (Ord. ANN model): 0.15 - Average Mean Absolute Error (AMAE) (KDLOR): 1.21

Haydon GH	2005	Self Organizing Maps (SOM - a type of ANN)	Unsupervised Learning	Not mentioned	Not mentioned	Not mentioned	Not mentioned
Hoot N.	2005	Bayesian Network	Supervised Learning	Several models based on Cox PH	AUC ROC between 0.6 and 0.7	Not mentioned	3-fold cross-validation on training set - AUC ROC: 0.67 Test set - AUC ROC: 0.68
Lau L	2017	Random forests ANNs	Supervised Learning	Logistic Regression	Not mentioned	AUC ROC - DRI: 0.68 - SOFT: 0.64	Best ML - AUC ROC (ANN): 0.84
Perez-Ortiz M	2017	Semi-supervised (Combined labelled and unlabelled data) - SVM variants Supervised - SVM variants - ANNs fitted using evolutionary algorithms (MPENGSA2)	Combination of supervised and unsupervised learning (semi-supervised)	Not mentioned	Not mentioned	Not mentioned	Best ML (3 month) - Accuracy (SVC): 0.90 - Geometric mean of sensitivities (GMS): (SVC-LP) 50.35 Best ML (12 month) - Accuracy (CS-SVC): 0.90 - Geometric mean of sensitivities (GMS) (CS-SVC): 55.09
Zhang M	2012	ANN	Supervised Learning	Not mentioned	Not mentioned	Not mentioned	C-Index - 1 year: 0.91 - 2 year: 0.89 - 5 year: 0.84

Table 3 - Results of Machine Learning methodology compared to regression and liver risk index scoring

Author	Year	Patient number	Standard LT score model used	Follow up duration (i.e. graft survival)	Number/fraction of graft failures	Number of input variables used (i.e. donor and recipient characteristics)	Validation data split and methods
Briceno J.	2014	1,003	Current validated scores (MELD, D-MELD [16], DRI, P-SOFT, SOFT, and BAR) - for transplant.	3 months	Not mentioned	26 recipients, 19 donor, 6 transplants	10-fold stratified cross-validation
Cruz-Ramírez M	2013	1,003	MELD (not quantitatively compared)	3 months	~100 (10%)	16 recipients, 20 donor, 3 transplants	75%/25% train-test split, with multiple bootstrap samples created from each set
Cruz-Ramírez M	2012	1,001	MELD (not quantitatively compared)	1 year	161 (16%)	16 recipients, 16 donor, 3 transplants	5-fold stratified cross-validation
Dorado-Moreno M	2017	1,406	MELD	Ordinal failure time categories: 0 - 15 days 15 days - 3 months 3 months - 1 year 1 year +	15% within 1 year	16 recipients, 17 donors, 5 transplants	10-fold stratified cross-validation Over-sampling of synthetic D-R pairs also used to increase the frequency of minority classes
Haydon GH	2005	827	MELD (not quantitatively compared)	3 months 1 year	Not mentioned	37 recipients, 18 donors	Separate dataset of 2,622 patients used for validation
Hoot N.	2005	12,239	MELD	3 months	13.4% (training set), 10.8% (validation set)	29 variables	2000-2001 data used for training. 2002 data used for validation. 3-fold cross validation also performed on training set.
Lau L	2017	180	DRI SOFT	1 month	11 (6.1%)	173 recipients, 103 donor initially then top 15 following feature selection	1,000 bootstrap samples with out-of-bag samples used for validation
Perez-Ortiz M	2017	822	MELD	3 months 1 year	Not mentioned	16 recipients, 17 donor, 4 transplants	Stratified 10-fold cross-validation. Unlabelled data from recent transplants and virtual DR-pairs also incorporated in varying quantities
Zhang M	2012	290	MELD and MELD-Na	1 year 2 year 5 year	119 (41%) patients died within 5 years	12 recipients, 2 donors (following forward stepwise feature selection)	80%/20% train-test split with additional 20% validation set created from training set.

Table 4: Study input variable and validation methodology

References

1. Neuberger J. Liver transplantation in the United Kingdom. Liver Transpl 2016;22:1129-1135.
2. Kim WR, Lake JR, Smith JM, Skeans MA, Schladt DP, Edwards EB, et al. OPTN/SRTR 2013 annual data report: liver. Am J Transplant 2015;15(suppl 2):1-28.
3. Busuttil RW, Tanaka K. The utility of marginal donors in liver transplantation. Liver Transpl 2003;9:651-663.
4. Tector AJ, Mangus RS, Chestovich P, Vianna R, Fridell JA, Milgrom ML, et al. Use of extended criteria livers decreases wait time for liver transplantation without adversely impacting post- transplant survival. Ann Surg 2006;244:439-450.
5. Halldorson JB, Bakthavatsalam R, Fix O, Reyes JD, Perkins JD. D-MELD, a simple predictor of post liver transplant mortality for optimization of donor/recipient matching. Am J Transplant 2009;9:318-326.
6. Croome KP, Marotta P,Wall WJ,Dale C,Levstik MA, Chandok N, Hernandez-Alejandro R. Should a lower quality organ go to the least sick patient? Model for End-Stage Liver Disease score and donor risk index as predictors of early allograft dysfunction. Transplant Proc 2012;44:1303-1306.
7. Flores A, Asrani SK. The donor risk index: a decade of experience. Liver Transpl 2017;23:1216-1225.
8. Yeh H, Smoot E, Schoenfeld DA, Markmann JF. Geographic inequity in access to livers for transplantation. Transplantation 2011;91:479-486.
9. Lewsey JD, Dawwas M, Copley LP, Gimson A, Van der Meulen JH. Developing a prognostic model for 90-day mortality after liver transplantation based on pretransplant recipient factors. Transplantation 2006;82:898-907.
10. Collett D, Friend PJ, Watson CJ. Factors associated with short- and long-term liver graft survival in the United Kingdom: development of a UK donor liver index. Transplantation 2017;101:786-792.
11. Braat AE, Blok JJ, Putter H, Adam R, Burroughs AK, Rahmel AO, et al.; for European Liver and Intestine Transplant Association (ELITA) and Eurotransplant Liver Intestine Advisory Committee (ELIAC). The Eurotransplant donor risk index in liver transplantation: ET-DRI. Am J Transplant 2012;12:2789-2796.
12. Feng S, Goodrich NP, Bragg-Gresham JL, Dykstra DM, Punch JD, DebRoy MA, et al. Characteristics associated with liver graft failure: the concept of a donor risk index. Am J Transplant 2006;6:783-790.
13. Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature 2015;521:452-459.
14. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255-260.
15. Deo RC. Machine learning in medicine. Circulation 2015;132:1920-1930.
16. Murphy K. Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press; 2012.
17. Cox DR. Regression models and life tables (with discussion). J Roy Stat Soc: Ser B 1972;34:187-220.
18. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019;28:73-81.
19. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li S-X, et al. Analysis of machine learning tech- niques for heart failure readmissions. Circ Cardiovasc Qual Outcomes 2016;9:629-640.
20. Critical Appraisal Skills Programme. CASP Cohort Checklist. 2018. https://casp-uk.net/wp-content/uploads/2018/01/CASP-Cohort-Study-Checklist_2018.pdf. Accessed October 1, 2019.
21. Briceño J, Cruz-Ramírez M, Prieto M, Navasa M, Ortiz de Urbina J, Orti R, et al. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: results from a multicenter Spanish study. J Hepatol 2014;61:1020-1028.

22. Cruz-Ramírez M, Hervás-Martínez C, Fernández C, Briceño J, de la Mata M. Multi-objective evolutionary algorithm for donor-recipient decision system in liver transplants. Eur J Oper Res 2012;222:317-327.
23. Cruz-Ramírez M, Hervás-Martínez C, Fernández JC, Briceño J, de la Mata M. Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. Artif Intell Med 2013;58:37-49.
24. Dorado-Moreno M, Pérez-Ortiz M, Gutiérrez PA, Ciria R, Briceño J, Hervás-Martínez C. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver trans- plantation problem. Artif Intell Med 2017;77:1-11.
25. Haydon GH, Hiltunen Y, Lucey MR, Collett D, Gunson B, Murphy N, et al. Self-organizing maps can determine outcome and match recipients and donors at orthotopic liver transplantation. Transplantation 2005;79:213-218.
26. Hoot N, Aronsky D. Using Bayesian networks to predict survival of liver transplant patients. AMIA Annu Symp Proc 2005: 345-349.
27. Lau L, Kankanige Y, Rubinstein B, Jones R, Christophi C, Muralidharan V, Bailey J. Machine-learning algorithms predict graft failure after liver transplantation. Transplantation 2017;101:e125-e132.
28. Pérez-Ortiz M, Gutiérrez PA, Ayllón-Terán MD, Heaton N, Ciria R, Briceño J, Hervás-Martínez C. Synthetic semi- supervised learning in imbalanced domains: constructing a model for donor-recipient matching in liver transplantation. Knowledge- Based Syst 2017;2017:75-87.
29. Zhang M, YinF, Chen B, Li B, Li YP, Yan LN, Wen TF. Mortality risk after liver transplantation in hepatocellular carcinoma recipients: a nonlinear predictive model. Surgery 2012;151: 889-897.
30. Tusch G. An optimization model for sequential decision-making applied to risk prediction after liver resection and transplantation. Proc AMIA Symp 1999:425-429.
31. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928-935.
32. Sousa FS, Hummel AD, Maciel RF, Cohrs FM, Falcão AEJ, Teixeira F, et al. Application of the intelligent techniques in transplantation databases: a review of articles published in 2009 and 2010. Transplant Proc 2011;43:1340-1342.
33. Senanayake S, White N, Graves N, Healy H, Baboolal K, Kularatna S. Machine learning in predicting graft failure following kidney transplantation: a systematic review of published predictive models. Int J Med Inform 2019;130:103957.
34. Nursetyo AA, Syed-Abdul S, Uddin M, Li YJ. Graft rejection prediction following kidney transplantation using machine learning techniques: a systematic review and meta-analysis. Stud Health Technol Inform 2019;21:10-14.
35. NHS Blood and Transplant Organ Donation and Transplantation Directorate Advisory Group Chairs Committee. Liver National Allocation Scheme. Fixed Term Working Unit—Organ Allocation. Bristol, UK: National Health Service; 2014.
36. Baxt WG. Application of artificial neural network to clinical medicine. Lancet 1995;346:1135-1138.