frontiers | Frontiers in Immunology

TYPE Review
PUBLISHED 11 November 2022
DOI 10.3389/fimmu.2022.966329

Artificial intelligence for precision medicine in autoimmune liver disease

Alessio Gerussi[1,2*], Miki Scaravaglio[1,2], Laura Cristoferi[1,2,3], Damiano Verda[4], Chiara Milani[1,2], Elisabetta De Bernardi[5], Davide Ippolito[6], Rosanna Asselta[7,8], Pietro Invernizzi[1,2], Jakob Nikolas Kather[9,10] and Marco Carbone[1,2*]

[1]Division of Gastroenterology, Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy, [2]European Reference Network on Hepatological Diseases (ERN RARE-LIVER), San Gerardo Hospital, Monza, Italy, [3]Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy, [4]Rulex Inc., Newton, MA, United States, [5]Department of Medicine and Surgery and Tecnomed Foundation, University of Milano - Bicocca, Monza, Italy, [6]Department of Radiology, San Gerardo Hospital, Monza, Italy, [7]Humanitas Clinical and Research Center, Rozzano, Milan, Italy, [8]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy, [9]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany, [10]Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany

OPEN ACCESS

EDITED BY
Jochen Mattner,
University of Erlangen Nuremberg,
Germany

REVIEWED BY
Bastian Engel,
Hannover Medical School, Germany
Mohammed Abu El-Hamd,
Sohag University, Egypt
David Assis,
Yale University, United States
Jie Wu,
Segmed, Inc., United States

*CORRESPONDENCE
Marco Carbone
marco.carbone@unimib.it
Alessio Gerussi
alessio.gerussi@unimib.it

SPECIALTY SECTION
This article was submitted to
Vaccines and Molecular Therapeutics,
a section of the journal
Frontiers in Immunology

RECEIVED 10 June 2022
ACCEPTED 13 October 2022
PUBLISHED 11 November 2022

CITATION
Gerussi A, Scaravaglio M, Cristoferi L, Verda D, Milani C, De Bernardi E, Ippolito D, Asselta R, Invernizzi P, Kather JN and Carbone M (2022) Artificial intelligence for precision medicine in autoimmune liver disease. *Front. Immunol.* 13:966329. doi: 10.3389/fimmu.2022.966329

Autoimmune liver diseases (AiLDs) are rare autoimmune conditions of the liver and the biliary tree with unknown etiology and limited treatment options. AiLDs are inherently characterized by a high degree of complexity, which poses great challenges in understanding their etiopathogenesis, developing novel biomarkers and risk-stratification tools, and, eventually, generating new drugs. Artificial intelligence (AI) is considered one of the best candidates to support researchers and clinicians in making sense of biological complexity. In this review, we offer a primer on AI and machine learning for clinicians, and discuss recent available literature on its applications in medicine and more specifically how it can help to tackle major unmet needs in AiLDs.

KEYWORDS

machine learning, genomics, radiomics, digital pathology, autoimmunity, deep learning, whole-slide digital image analysis, population genetics

1 Introduction

Autoimmune liver diseases (AiLDs) are chronic diseases affecting the liver and the biliary tract, with a putative autoimmune pathogenesis, and include autoimmune hepatitis (AIH) (1), primary biliary cholangitis (PBC) (2), and primary sclerosing cholangitis (PSC) (3). The combination of low prevalence, unknown etiology, and high degree of heterogeneity among patients fulfilling the same diagnostic criteria have hitherto hindered the development of drugs, especially for PSC and AIH.

Nonetheless, high-throughput DNA and RNA sequencing technologies, digital pathology, and digital radiology are also progressively reaching this neglected field. A large amount of experimental and clinical data are increasingly available in the field (4), which requires dedicated analytical pipelines that are able to deal with big data. Artificial intelligence (AI) is a broad scientific field including many sub-specialties. Figure 1A summarizes the relationship between AI, machine learning (ML), and deep learning (DL). AI comprises several sub-fields, and the most important ones for medical applications are ML and DL. ML algorithms create models that learn from sample data (training data) and are then able to make inference/predictions on new data without being explicitly programmed for this scope. Among the others, ML is of particular interest for the biomedical field, since it can recognize patterns within data and leverage them to generate new biological knowledge. DL is a sub-field of ML that uses multiple layers of information for extraction of features from raw inputs. This type of AI is particularly well-suited for image processing.

This review aims to provide an introduction to AI for clinicians and to outline the current evidence about AI in medicine and the foreseeable applications in the field of AiLDs.

2 Artificial intelligence: Working definitions and its growing role in the biomedical field

In 1959, Artur L. Samuel, a computer scientist, firstly introduced the expression “machine learning” in his seminal paper focusing on how a machine could learn the game of checkers (5). In line with Samuel’s definition, ML could be defined as that sub-field of AI in which computers are not explicitly programmed by experts but rather learn from experience (6). For instance, this could be the case of a process that analyzes historical data concerning the time needed to recover from a disease, and it is later used to predict the recovery time for new patients with the same disease. In this example, a set of *labeled* data is available, e.g., the number of days needed to recover; this value is the value of the *target* (alternatively referred to as the *output*) of the problem at hand, and it is known for a set of the patients tracked in the historical record. When the value of the output is known and labeled, the problem falls into the area of *supervised learning*. When the output of interest is a *quantity* (e.g., the number of days needed to recover), the problem is referred to as a *regression*. On the contrary, if the *output* represents a *quality* (a binary one: disease status versus healthy status), the problem would be defined as a *classification*.

In other cases, the set of *labeled* data is not available and the macro-category is called *unsupervised learning*. For example, unsupervised learning may refer to the task of splitting a set of patients into homogeneous subgroups with respect to a set of features (*clustering*) or rather to determine which diagnostic factors are correlated to each other (*association mining)*. In these cases, there is no known, pre-defined and pre-labeled target of interest.

All these tasks can potentially be addressed with a statistical approach. The main difference between statistics and ML is that ML does not require to make any assumptions concerning the statistical distribution of the considered features (7). Conversely, the outcome of a statistical pipeline relies on (and benefits from) the knowledge about the underlying distribution of the considered population, as well as the statistical properties of the chosen estimator. This requirement makes the statistical approach less affordable in the case of high-dimensional data, as it further increases complexity.

Conversely, considering that ML approaches cannot be compared against any reference distribution for evaluation, it is harder to assess their performance. A common procedure to

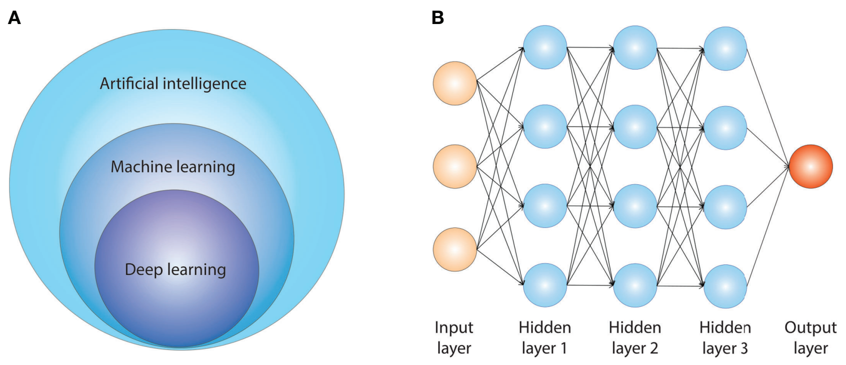

FIGURE 1
Artificial intelligence and its sub-fields.

deal with this issue consists in comparing the model predictions against the data themselves. More specifically, a subset of data (for instance, a subset of patients whose data are available in the historical record) is used to *train* the model, while another part is not supplied to the training pipeline. In ML terminology, the former part is usually referred to as *training set*, while the latter is commonly denoted as *test set*. After having completed the modeling phase on the *training set*, the model itself is applied on the *test set*; in other words, the model will make predictions on a new and unknown subset of data, which simulates the foreseeable new data on which it will be applied in the future. Coming back to the first example about days to recover from a disease, the *test set* will involve patients not considered in the modeling phase on training. More complex and sophisticated scenarios also involve the introduction of a third dataset, referred to as the *validation* dataset; in this case, there is a dataset for *training*, a dataset for *hyperparameter tuning* (tuning of model parameters) on top of training, and another one for *performance evaluation*.

External validation should be performed for rigorous evaluation of the algorithm performance. In fact, deficiency and biases present in the training dataset may appear in external data that are either closer to the ideal target population or representative of minority populations. In this way, the risk of biased performance estimation is reduced and more consistent and robust conclusions can be drawn (8).

Evaluating, for instance, the *root mean squared error* (or, in the case of a *classification*, other indicators such as *empirical accuracy* or *area under the curve*) allows to understand how much the extracted model is effective on previously unseen data. When there is a strikingly good performance in the *training set*, but with a poor one in the *test set*, the model is *overfitting* data: in other words, it is too focused on replicating training data and is not able to generalize its predictive power.

Among the different ML available models, some of the well-known approaches are artificial neural networks (9) and support vector machines (10). In the latest couple of decades, neural networks and especially multi-layered ones became even more popular, being at the core of DL approaches. DL refers to a pipeline in which the learning process is modularized: the first layer of modeling can be considered as in charge of learning features that will be used by the following layer, enabling the final one to provide a prediction (11). This has been shown to be particularly promising in the field of high-dimensional, *unstructured data*, such as documents or images (12). Figure 1B shows the basic architecture of an artificial neural network. DL algorithms have a more sophisticated and complex architecture than non-DL ML algorithms, which include many more free parameters, providing a higher degree of flexibility. The core of DL algorithms is usually the artificial neural network, which is constituted of several nodes called artificial neurons. The output of a neuron is a non-linear computation of the sum of its inputs.

Overall, AI and its sub-fields ML and DL offer a wide range of tools for data mining and predictions, which represent a great opportunity to advance the biomedical field. Multiple examples of AI applications have been produced over the last years in several fields, such as pathology, radiology, dermatology, and endoscopy, to name a few (13). Analyses of images derived from tissue specimens (14), radiological exams (15), and pictures of skin lesions (16) or captured from videos of endoscopic procedures (17) are examples of input data that have been analyzed through different types of neural network algorithms. In addition, clinical data either extracted from multicenter collaborative efforts or derived from electronic health records (EHRs) have also been analyzed with ML software to understand whether the AI-based models were better than available diagnostic and/or prognostic scores (18) or to identify subgroups of individuals at different disease course (19, 20). Neural networks have also shown promising results in several fields of genomics (21); particularly interesting are also approaches that integrate genomic data (most commonly common variants from genome-wide association studies) with images, with the aim to match image-derived markers with gene signatures (15).

3 Applications in autoimmune liver diseases

3.1 Digital and computational pathology

3.1.1 Aims and applications

Histopathology slides intrinsically hold a large amount of information and data, which have been largely underutilized in the past. The transformation of these analogic data in digital file formats is the core of “digital pathology” (22). The whole slide imaging (WSI) technique implies the digitalization of the whole histological section *via* a digital scanner, and has been progressively becoming more available and widespread.

The revolution carried forward by ML, particularly in its sub-domain of DL, is fostering the development of an associated new field called “computational pathology” (23). This sub-field regards all the processes involved in extracting and handling data present in digital slides to generate valuable information for clinical and research purposes. Overall, computational pathology may be capable to provide solutions to several issues in modern medicine. Schematically, we can enucleate several applications: in the diagnostic area, it favors automation, supports the pathologist, and enhances telemedicine; in the research area, it improves the understanding of the pathogenesis at the cellular and tissue levels, prognostication, and risk stratification.

In hepatology, and in rare disorders like AiLDs, histological evaluation of liver tissue may play an important role in the diagnostic workout. Moreover, it may offer semiquantitative/quantitative prognostic information and supports choice of therapies. Therefore, hepatology represents an appropriate field for WSI and computational pathology (24).

3.1.2 Whole slide imaging technology and deep learning

The first step in WSI technology is the use of the appropriate technology for image acquisition, i.e., an image scanner for digital image acquisition. The second part of the process is based on the view and analysis of the image through a dedicated software (25).

The most common method to generate images is tiling, i.e., acquiring the original slide as tiles, although a linear scanning system can also be used. The final image is the result of merging of each tile or line scan by the software. DL algorithms extract parameters of interest from the scans by using labels corresponding to predefined categories. They can obviously be used also through an unsupervised manner for clustering or other data grouping strategies.

It is evident that WSI technology coupled with DL could work in a much more automated and efficient way than current standards, and can be of help for clinical practice (26). Arguments in favor of this statement are the improved reproducibility and speed of the diagnostic process together with the reduction of workload for clinicians (26). By reducing the need for repetitive work, pathologists can focus on those tasks that require specific skills like the integration of the available information to produce accurate diagnostic hypotheses and communicate with clinicians. Before translation into clinical practice, it is essential to validate WSI performance compared to standard procedures; few studies are available, but data seem promising (27). Most of the published studies in the hepatological field include only WSI without subsequent DL applications, and are mainly focused on exploring the added value of WSI in terms of better inter-observer concordance. In fact, the digitalization allows better sharing and annotation of tissue slides, making easier collaboration among pathologists (22).

Among the studies that went beyond pure assessment of reproducibility of the diagnostic process, a recent and elegant work from Cheng et al. has shown the excellent accuracy of a DL algorithm applied to images derived from formalin-fixed, paraffin-embedded surgical resections and biopsy specimens of nodular lesions of the liver. On WSI, a pathologist outlined ROIs generating hundreds of thousands patches for further analysis. Several testing sets were available for independent external validation. Three DL models were evaluated, with AUC always > 0.90 in external validation datasets. Of note, to avoid batch effects and increase generalizability, data for external validation were collected from three different hospitals. To us, the core elements of this work are a solid methodology, the involvement of expert pathologists, and the use of a large training set together with heterogeneous validation sets taken from different centers (28).

Another important study from China has shown the capacity of DL applied to WSI derived from hepatocellular carcinoma (HCC) to discriminate the type of tissue and to detect new prognostic biomarkers. In this case, on top of data coming from a single Chinese cohort, the investigators have also leveraged the Cancer Genome Atlas (TCGA) to finely map the identified histological patterns and correlate them with tumor immune infiltrates and gene mutations. Of note, to take into account the different staining protocols between the training cohort and TCGA, they used a specific DL algorithm for standardization. This work represents a remarkable example of a pipeline that goes from slice preparation to the development of a risk score associated to histological patterns identified within the tumor, showing the ability of this score to stratify patients in higher- and lower-risk groups and to correlate score values with histological patterns and tumor-associated immune cells (29).

More relevant to the field of AiLD are applications of WSI and DL to inflammatory diseases such as ulcerative colitis (UC). In a recent landmark paper showing data from an international multicenter consortium, investigators have created a new histological score for UC, which correlates with endoscopic findings and holds prognostic value. To generate the score, colonic biopsies were used and a group of expert pathologists annotated the slides by using a variety of histological scoring schemes already available. After a Delphi consensus among pathologists, the neutrophil infiltrate was deemed as the key element of disease activity and clinical outcome; importantly, the new score was created by using standard statistical methods. Subsequently, the DL algorithm was trained to learn how to identify neutrophil infiltrates within images and to differentiate between quiescent and active disease, achieving 86% accuracy. In our opinion, the interesting aspect of this study is the blended approach between classical and standardized statistical approaches (development of scores and survival analysis) together with DL (17).

WSI offers a more quantitative approach for assessment of liver fibrosis and steatosis, and novel data are available for HCC and transplant pathology (22). A dedicated review of the most recent applications of WSI technology to liver diseases is reported below (22). No specific data are available for AiLD at the time of writing.

3.1.3 Methodological hurdles and limitations

The main limitations to the implementation of AI-based technologies to digital pathology are the lack of universal standards for data formatting, as compared to radiology, where the Digital Imaging and Communications in Medicine (DICOM) format is already the standard (6). The current trend in the field is to use AI to digitalize these processes and reduce arbitrariness and low agreement among pathologists (30).

In addition, data quality is essential; in fact, histological slides comprise a highly heterogeneous information. Staining, thickness of the section, and presence of artifacts are known factors influencing the model performance in DL-based diagnostic models (31). The development and optimization of staining methods to enhance the contrast of biological components has been a goal for decades. Yet, inconsistencies and artifacts are still generated despite technical efforts to improve specimen preparation. This issue represents an obstacle for digital pathology, since a batch effect can be introduced when analyzing altogether samples either from different institutions or from the same institution but retrieved at different time points (32). Novel strategies have been put in practice to overcome this hurdle by means of unsupervised methods based on color normalization and adversarial adaptation (33). Convolutional neural networks (a class of artificial neural network commonly applied for image analysis) are used to learn properties present in the source domain and then apply them on the target domain without any supervised labeling (33, 34). Several specific strategies for prevention of the artifact-driven loss of performance are currently under investigation (31).

Another important limitation of current pipelines that is particularly relevant for the field of AiLD is their exclusive reliance on hematoxylin–eosin stainings. Despite the validated role of CK7 and orcein stainings in the differential diagnosis and staging of PBC (35) and PSC (36), to our knowledge, there are no published computational pathology pipelines that have been trained on this type of images.

Overall, there are still some methodological limitations for the full implementation of digital pathology as a research tool and for its incorporation in the current clinical workflow; yet, there is active research aiming at addressing them.

3.1.4 Potential applications in AiLD

Liver biopsy has a pivotal role in the diagnostic and prognostic process of AiLDs. It is essential for the diagnosis of AIH and holds value in atypical cases of PBC and PSC; for all the three conditions, it provides a rich amount of information that can assist prognostication and guide treatment. While for PBC the pediatric onset is exceptional, AIH and PSC can arise at pediatric age, with features and disease course different to the adult onset.

Computational pathology has several promising applications in the field of AiLDs. As the pathogenesis of these diseases is still obscure, the use of AI-assisted methods may aid in discovering pathogenetic clues for further investigation. Computational pathology may also assist in identifying core histological features of AiLDs that are still missing, and provide a more standardized approach for differential diagnosis (e.g., discriminating between pure PBC or AIH and variant syndromes). An even more promising application is for prognosis modeling and risk stratification.

Nevertheless, there are some disease-specific traits that can influence AI implementation. In AIH, liver biopsy is the cornerstone of the diagnosis, and the available scoring systems unlikely lead to definite AIH diagnosis without histological evaluation. In addition, histology provides useful information on disease activity and stage (37). Histological diagnostic criteria evolve over time, and AIH makes no exception. The combination of interface hepatitis and a predominantly periportal lympho-plasmacellular infiltrate is nearly always present in AIH, but it is not pathognomonic (38). The recent revision of histological criteria for the diagnosis of AIH has recognized centrilobular injury, together with central perivenulitis and necrosis, as being part of the histological spectrum of acute severe AIH (39). However, qualitative or semi-quantitative information derived from liver histology of patients with AIH is currently insufficient to accurately depict its heterogeneous histological phenotypes and to predict treatment response and/or relapse. Moreover, the differential diagnosis with other forms of hepatitis, especially acute hepatitis or drug-induced liver injury, remains extremely difficult. A foreseeable goal would be to have quantitative metrics of subtle processes such as the extension of lympho-plasmacellular infiltrates within the periportal tract, lobular necrosis, and other pathological processes that are frequently observed in AIH, to develop a more reliable diagnostic and prognosis prediction approach. It appears evident that AIH would hugely benefit from the application of digital pathology because the digital analysis of histological slides could offer a standardized approach to its diagnosis and could help to identify features that are inherently proper of AIH rather than drug-induced liver injury or other AIH-mimics (40). ML is required to deal with the large amount of information that would be obtained after extraction of quantitative metrics from digital slides (feature selection).

For risk stratification in AIH, unsupervised ML approaches could be of interest to identify prognostic biomarkers of disease activity that can predict biochemical remission, when liver biopsy is performed at diagnosis, or relapse, if the histological assessment is performed before treatment withdrawal. Unsupervised learning could shed a light on morphological features peculiar of sub-phenotypes of AIH that are currently invisible to the human eye, highlighting populations of cells or morphological patterns that are not considered part of the histological spectrum of the disease (41–44). Unsupervised learning techniques are commonly used to analyze and interpret single-cell RNA sequencing data; for instance, Liu et al. have recently shown that single-cell profiling of immune transcriptomes of skin samples from patients with different inflammatory skin disorders can differentiate among different conditions by identifying, in a unbiased manner, gene expression signatures (41). On a similar note, the analysis of approximately 200 million nuclei from digitized slides of 117 patients affected by glioblastoma was able to derive three disease clusters with different nuclei morphology and specific associations with gene signatures (44).

Even though only supervised learning approaches were employed, in another landmark paper from Stanford University, it was shown that the use of a computational pathology tool that automatically generates quantitative features can pinpoint structures that had not been called in action as prognostic biomarkers. This is a clear example of how the implementation of quantitative approaches, either supervised or unsupervised, can represent a way to detect unseen patterns that hold predictive and prognostic value (43).

Multi-center, collaborative efforts can represent an asset by helping the collection of a huge amount of clinical and histological data to be analyzed through ML; large consortia such as European Reference Network for Rare Liver Diseases (ERN-RARE LIVER) or the International AIH Group can be leveraged to this end. Unsupervised techniques can be devised and validated in order to cluster individuals affected by AIH according to different prognostic trajectories.

As regards autoimmune cholangiopathies, histological samples suitable for analyses are limited in number. The diagnostic accuracy of PBC-specific autoantibodies has determined a progressive reduction in the number of liver biopsies performed (45); in PSC, liver biopsy is required only in atypical cases, since diagnosis and monitoring are performed by magnetic resonance cholangiopancreatography (MRCP) (46). Notwithstanding, PBC and PSC could potentially benefit from WSI together with ML. PBC and PSC are rare diseases with poorly understood pathogenesis; quantitative analysis of histological slides compared to healthy controls and/or other liver diseases such as metabolic liver disease or viral hepatitis could point toward zonal and cellular differences that are specific for these conditions. In this way, highlighting histological areas of potential interest, ML can be a tool to facilitate hypothesis generation of novel physiopathological models that can be subsequently dissected in the lab. Furthermore, the implementation of AI in the histological assessment of autoimmune cholangiopathies may be of aid in differential diagnosis with other chronic cholestatic syndromes. Of note, the definition of small duct PSC is still problematic and its distinction from intrahepatic genetic cholestasis or autoantibody-negative PBC is a challenge. Whether quantitative information included in histological slides of patients with small duct PSC is useful to predict the evolution toward a large duct form is still unknown, and it might be worth exploring in an integrated approach together with AI applications to radiology.

As regards biomarker discovery and risk stratification, non-invasive tools are not accurate enough to depict the full cholestatic picture, and the variety of inflammatory and fibrotic patterns, together with the cellular milieu of biliary regeneration, are not captured by transient elastography or routine liver enzymes (47, 48). This is not trivial; there is increasing evidence of the prognostic role of ductular reaction in these conditions (47). The alterations of the physiological architecture of the liver and biliary tree architecture pinpointed by ML could also represent novel biomarkers. Quantitative biomarkers of inflammation, biliary damage, or fibrosis can be discovered by means of AI and correlated with non-invasive biomarkers.

As regards disease classification and definition, unsupervised learning algorithms can be applied to detect sub-phenotypes that could prompt disease re-classification. The latter aspect is of particular interest for variant syndromes and for better characterization of the ductopenic variant of PBC.

Overall, the introduction of ML in this field of medicine may at minimum generate new hypotheses and identify novel biomarkers; yet, robust studies validated in several cohorts will be required to change also everyday clinical practice.

3.2 Applications in radiology

3.2.1 Aims and applications

In the last decade, the remarkable progress in liver imaging techniques has helped to characterize several liver diseases from a qualitative and quantitative point of view. The foreseeable introduction of AI in medical practice together with the generation of a large amount of high-quality imaging data poses radiology as a key player in precision medicine (13, 49).

To this end, two groups of AI-based techniques can be mentioned: radiomics, which relies on ML, and DL systems (based on neural networks) (6). Radiomics has emerged as a high-throughput computing technique that enables extraction of large amounts of quantitative features from medical imaging, mainly computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) (50). This vast amount of variables can be correlated with specific clinical outcomes of interest, providing far more information than those detectable by an experienced physician (51). The main difference between radiomics and DL (also known as deep radiomics) is the methodology of feature extraction. In "standard" radiomics, image analysis experts derive a list of mathematical equations that are applied to the image; in DL, convolutional neural networks are used for automatic extraction of features without the need of pre-defined programming (30).

We can speculate that the ultimate goal of both techniques is the combination of radiological data with clinical and laboratory data and potentially other -omics, to develop more accurate predictive models that incorporate a wider spectrum of disease-related features (52, 53).

3.2.2 Methodology

The methodological process has been classically divided in distinct phases (51). The prerequisite is the collection of high-quality images with standardized imaging protocols to allow the repeatability and reproducibility of the analysis. In case of multicenter studies, as image resolution and intensity can be different depending on image acquisition and reconstruction procedure, preprocessing of the collected images is mandatory. This step is typically called *data acquisition and normalization*.

After preprocessing, one should define a region of interest (*segmentation of region of interest or ROI*). In most radiomics studies on liver diseases, the segmentation is performed by a radiologist as manual segmentation. Alternatively, the identification of ROI can be done by computer analysis through specific algorithms (automatic segmentation), with an optional input provided by the radiologist (semiautomatic segmentation). From the defined ROI, quantitative data are subsequently extracted (*feature extraction*). Informative data include both manual engineered features, like patterns of intensity, texture and shape, and abstract DL features. Of all the quantitative features extracted from the ROI, only the most informative will be retained, after feature selection through computational methods. Selected features are then fitted in a specific model of analysis. While a single modeling technique is often used, multiple-modeling methodology must be preferred to limit effects on prediction performance. Internal and possibly external validation of the model should be performed to avoid overfitting the model (*features selection, modeling, and validation*). The last and most important step of radiomics is to correlate the selected characteristics with the outcomes of interest to better characterize the disease (*image analysis*). Recently, the radiomics quality score (RQS) has been proposed to evaluate if a radiomic study matches with defined quality criteria in all steps (50, 51).

Overall, there are several methodological steps, and each of them needs high level of control to avoid biases in the prediction.

3.2.3 Limitations

The first obstacle to the widespread application of radiomics in the study of liver diseases is the use of non-standardized image acquisition and reconstruction protocols even within the same institution, together with the need of a large amount of data, which is challenging and time-consuming. Moving forward through the radiomics workflow, the segmentation process has also some limitations. While manual segmentation is time-consuming, many automated and semi-automated algorithms are often suboptimal so that physicians are almost always needed to verify their accuracy. Moreover, in rare diseases such as AiLDs, automated segmentation algorithms do not exist. One solution might be the extraction of features through neural networks (54); the downside of this approach for AiLDs is the rarity of these conditions, while DL typically requires large datasets for training.

Another important issue that limits the broader adoption of AI algorithms is the lack of interpretability, which is the *black box* problem, namely, the difficulty of physicians to understand the predictions of ML algorithms (55) (see also *Section 3.5* for the debate about black box and explainable AI).

Overall, similarly to digital pathology, radiomics can also suffer from lack of standardization. It is conceivable that some strategies to address this issue can be shared between the two fields but most tasks are field-specific.

3.2.4 Applications in AiLDs

Imaging plays a remarkable role in the diagnosis and management of PSC. Among medical imaging techniques, MRCP represents the main non-invasive imaging method for the diagnosis, risk stratification, and monitoring of patients with PSC (56). Yet, there is still lack of radiological features specific for PSC that allow exclusion of other causes of cholangiopathy. Thus, there is particular interest in the potential of quantitative imaging in terms of phenotypic characterization and differential diagnosis.

The highly variable disease course of PSC, likely associated with a variety of uncharacterized sub-phenotypes, represents a challenge for risk stratification. Several attempts have been made to develop reliable predictive tools in order to early discriminate patients with a more aggressive disease (56). To date, prognostic scores have been based mostly on laboratory and clinical data; some of them have been created by the application of ML algorithms, but without the inclusion of radiomic features (57, 58). Unfortunately, the fluctuating nature of serum markers of cholestasis during the course of the disease has hampered the accuracy of models based only on laboratory values so far. Liver stiffness measurements (LSMs) by either transient elastography or MRI elastography hold prognostic value (59, 60). Early arterial peribiliary hyperenhancement at MRI has been associated with higher Mayo risk scores and poorer prognosis (61). The ultrasound evaluation of incremented spleen size has also been correlated with major clinical outcomes in PSC (62).

The ultimate turning point in the evolution of radiological PSC characterization is the possibility to derive quantitative data from MRCP scans. There is mounting evidence on the accuracy of MRCP+, a novel image processing software that is able, through the creation of a 3D-enhanced model of the biliary tree, to provide quantitative metrics of the ductal anatomy, generating data on biliary tree volume, median diameter of the extrahepatic bile ducts, and number, length, and severity of strictures and dilatations (63). AI takes part in the segmenting, enhancing, and pre-processing of images together with the modeling of the derived information.

There is evidence supporting the reliability of MRCP+ metrics as a non-invasive tool to differentiate pediatric PSC from pediatric AIH (64, 65). In addition, MRCP+ parameters hold prognostic value, as proven by their strong correlation with validated biochemical and semi-quantitative MRCP-based risk scoring systems (66, 67). The application of AI on large-scale

MRCP+-derived quantitative data has the potential to significantly improve the current diagnostic approach and prediction models of PSC.

As regards AIH, promising data employing Liver MultiScan technology have been recently presented. Following the evidence that multiparametric MRI (mpMRI) using iron-corrected T1 (cT1) relaxation maps provides an accurate, non-invasive quantitative biomarker of liver fibrosis and inflammation, recent works have shown that mpMRI, when applied to AIH patients, has a better performance in detecting residual disease activity than serological biomarkers (68, 69). Moreover, it seems that higher cT1 value at diagnosis correlates with a higher risk of loss of biochemical remission, gaining prognostic value (68, 69). Liver MultiScan technology seems to be a foreseeable accurate non-invasive biomarker in AIH that can enhance risk stratification. Multicenter prospective studies are needed to validate these preliminary findings, together with the implementation of AI platforms to leverage the large amount of data generated by these new technologies.

In conclusion, future implementations of radiomics and DL systems have the potential to improve our comprehension of the complexity of AiLDs. Looking forward, there are several potential future developments. Novel biomarkers can be investigated and validated; they may play a role as diagnostic, prognostic, and predictive tools to assist current clinical practice and clinical trials. Correlations of radiomic features with molecular parameters can enhance their contribution and potentially shed light on novel disease sub-phenotypes.

3.3 Population genetics

3.3.1 Predicting phenotype based on genotype: A supervised learning approach

From a genetic perspective, AiLD are complex traits (70); in other words, their genetic architecture is not monogenic but dependent on the interplay of several genetic variants. The field of AiLD is still at the dawn of the big data era. While GWAS have been performed for each of the three conditions, whole-exome and whole-genome sequencing data are still missing. For all three conditions, single-nucleotide polymorphisms (SNPs) within the human leukocyte antigen region have a significant role in shaping their genetic risk (71–73); yet, several non-HLA variants have been described for PBC (74) and PSC (73), and more recently also for AIH (75). The discussion of the large topic of missing heritability is out of the scope of this review (76); however, it is worth mentioning that large portions of the heritability of AIH, PBC, and PSC are yet to be characterized (77). Whole-exome and whole-genome sequencing have clearly revealed that the identification of rare predisposing variants with large effect size is useful to fill this gap of knowledge (78, 79). Evidence in AiLD is scanty, mostly available for PSC, where autosomal-like patterns of inheritance have been identified in some families (80, 81), although it is likely that AiLDs derive in most cases from the interaction of some environmental triggers on the ground of a predisposing genetic background mostly composed of common variants. That said, the utility of PRSs, which are typically based on common variants, in AiLD is still a matter of debate (77).

In this paragraph, we focus our attention on the possible applications of ML on GWAS data, since they represent by far the largest data already available in this field.

ML is considered a complementary tool in population genetics, where several methodological hurdles need to be overcome. Research in population genetics has mostly focused on the formalization and validation of statistical models that describe patterns of variations and their application to experimental molecular data (82). While classical population genetics has been mainly characterized by parameter estimation in the context of a predetermined probabilistic model (typically the Wright–Fisher model), the target of ML is optimization of the accuracy of predictions (82). PRS predictions are based on a linear parametric regression model, with strict assumptions like additive effects, independent effects, normal distribution of the data, and independence of observations (83). These assumptions are often not valid in complex diseases like AiLDs. For example, thanks to their non-linearity, ML algorithms allow to account for complex interactive effects between associated alleles (84). Another peculiar and powerful feature of ML is its capacity to handle thousands of dependent variables, each characterized by a massive amount of information; this ability is of interest in the genomics world, where increasing dimensionality of data is an issue (82).

In population genetics, the output could be represented by the status (case or control) or a continuous phenotype (such as the value of a blood biomarker of interest), and the features are the individual sample genotype data (83). Data feature selection is the key step to obtain an accurate ML model (84). There are a few methods (embedded methods and wrappers) useful to select only informative SNPs as potential predictors (83).

The research question should be clear: does one want to predict outputs or to interpret data? The generative approach builds a model for two classes in a supervised manner, while the discriminative approach focuses only on separating them *via* an unsupervised approach.

The main application of supervised ML in population genetics is to build a model to classify cases and controls based on SNPs. This approach has the research aim to leverage AI to create a feature ranking of the most significant genetic variants that are inherently specific for the disease of interest and to create a polygenic model that can complement PRS. Moreover, ML can incorporate other relevant information such as sex to create hybrid models that can risk stratify the genetic liability already at birth. A caveat that should be carefully considered is that only "pure" phenotypes should be included as cases, to avoid confounding.

Another possible application of supervised ML in AiLDs could be to identify novel predictive features (SNPs) associated with phenotype, possibly looking at biologically distinct sub-phenotypes of the disease (early *vs* advanced disease, onset at younger age *vs* older age, positivity for specific autoantibodies). In this way, a predictive model is generated, taking advantage of the different contribution of variables within the training genotype data (83). After the training phase, the models with the maximum predictive power are selected for validation. This stage is essential to avoid overfitting and is usually achieved by cross-validation (dividing the original dataset into a training set and a test set). Nonetheless, external replication is still required for the final validation of the model (83).

Unsupervised approaches may be used to cluster patients according to genotype data and investigate whether these novel groups have different clinical presentations, trajectories, and treatment responses. Based on the availability of other omics, clustering can be extended to genomics and transcriptomic data for example. After generating clusters with hypothesis-free means, it is mandatory to understand if they hold biological and clinical significance, to create a classification that is really meaningful for clinicians. Yet, datasets having a different set of omics for the same group of individuals are seldom available. In addition, it is worth mentioning that, despite the growing body of GWAS data available in public repositories, there is still a large fraction of data that is not publicly shared with the community of researchers. Privacy issues do exist and the matter of balancing privacy rights with the need of sharing knowledge for the sake of the scientific advancement is a hot topic in genetics (85).

3.4 Studying gene–gene interactions: A task for unsupervised learning

ML can also be helpful for studying epistasis, a so far neglected topic that may account for part of the missing heritability (86). Epistasis is difficult to study in humans compared to simpler animal models such as *Drosophila melanogaster*, and has many computational issues. Statistical definition of epistasis is that of interaction, the departure from a linear model describing how a number of predictors (x_i) predict the outcome (the phenotype y). y can be a quantitative measure (e.g., height) or a binary outcome (case *vs*. control), so that linear or logistic models should be used, respectively.

If a locus of interest does have an influence on the phenotype and this happens *via* an interaction with another locus, it turns out that a model that incorporates interaction may increase the power to detect the effect of the locus of interest to the phenotype. For example, if the locus of interest is A, it may be more interesting to compare a model where the effects of locus A and B, and their interactions are included in a model where all terms (either main or interaction) involving locus A are removed.

Yet, in GWAS, several loci of interest should be investigated. The simple way is an exhaustive search of all possible pairs of loci studying interactions for each couple (two-locus interaction) or replicating the three degrees of freedom test iteratively. It is evident that a multiple testing issue arises and corrections should be employed, but this leads to the detection of only huge epistatic effects (87). Computational power (e.g., computer cluster) is required, but these analyses are still feasible; the real point is the lack of scalability to higher-order interactions. Pre-filtered loci that show some degree of significance in terms of correlation with the phenotype may be reasonable, but it is flawed by the exclusion from the analysis of those loci that do not show association with the phenotype (88).

One strategy is to guide the analysis based on biological plausibility (e.g., based on known interactions at the protein level, such as interactions among transcription factors and their targets, or proteins belonging to the same biological pathway). Another strategy is to use ML, which does not demand a marginal effect in place. ML can overcome the obstacles encountered by traditional regression-based methods since it works without a prespecified model and explores different models to search for the most computationally efficient one, avoiding a comprehensive research (89). From a mathematical point of view, ML does test for associations allowing interactions rather than testing directly for the interaction *per se* (see above). Another advantage is that there is a long line of research in computer science for problems like feature selection and data mining. For example, the Random forest method builds a tree-fashioned model measuring the effect of each SNP both individually and through interactions, generating a feature ranking, as compared to the list of *p*-values provided by PLINK, i.e., the gold standard software used in GWAS studies (89, 90). Since ML is not exhaustive and adopts heuristics, cross-validation and external validation steps are essential to avoid overfitting of the training set; if the sample size is small, parametric methods may be better suited for the analyses than ML techniques (91). An important caveat should be mentioned: epistasis can also be seen from a functional point of view (functional epistasis) rather than a statistical one. This means that epistasis occurs in biology and may be present despite lack of signals from quantitative studies (87, 88).

Overall, the investigation of gene–gene interactions has represented a methodological challenge for long. There is high expectation that AI-based pipelines can solve at least some issues, even though it is likely that large amounts of data will be required to fully recapitulate the network of gene–gene interactions.

3.5 Integrative multi-omics

ML represents a potent tool for analyses of data derived from high-throughput sequencing. As for other scientific fields, the possible applications of ML can be (1) generation of models for classification; (2) clustering of individuals in groups; and (3) feature selection. In this section, we provide several examples of these different applications.

The field of AiLD is still at the dawn of the big data era. Gene expression, proteomic, and metabolomic data come mainly from peripheral blood and single cohorts. More specifically, a recent study has shown that ML can discriminate between AIH and healthy controls based on gene expression profiles (92). A similar situation exists for microbiota, despite the growing interest on this side (93, 94). Single-cell data are becoming available for healthy liver in mice and human and in fibrotic livers (95); HCC and cholangiocarcinoma are also under investigation with these novel techniques.

Despite having a broad range of applications, ML is mostly useful for large datasets, such as those derived from microarray and high-throughput sequencing studies. Available data are genomic (SNP, whole exome, and whole genome), epigenetic (DNA methylation and histone modifications), transcriptomic (coding and non-coding RNAs, single-cell or bulk sequencing), proteomic, and metabolomic, with many others emerging (96). There is increasing availability of these datasets, which are frequently independently analyzed from different groups with different techniques: a brilliant example of such datasets is the UK Biobank (97). Datasets including the same set of omics for all individuals included in the cohort are rare. Most of the available omics data repositories were created without the vision of future multi-omics integration but rather to host data that were derived from a specific technology at a specific time. While linking different datasets is feasible, this is still extremely difficult, if not impossible, at the individual level. Cloud-based platforms for hosting several omics data for multi-omics integration have been developed (e.g., https://opendata.lifebit.ai/) (98). Despite shared guidelines have been produced (the FAIR Data Principles), the field is still lagging behind (99).

Integrative multi-omics is a rapidly growing field within systems biology (96). Multi-omics integration is considered a cornerstone of precision medicine initiative (100). Putting together different types of information can be important for biomarker discovery and risk stratification as well as for pathogenesis and disease definitions. A promising example of this approach has been presented by Wainberg et al. (101) (Figure 2). The cohort under study was taken from the Arivale Scientific Wellness program, which included thousands of subjects undergoing several analyses: whole-genome sequencing or SNP microarray genotyping, and proteomic, metabolomic, and clinical laboratory measurements from 2015 to 2019. Authors generated polygenic risk scores (PRS) for 54 traits previously investigated through genome-wide association studies (GWAS) and investigated correlations between genetic risk scores and analytes, revealing that healthy subjects with high genetic risk show dysregulations of analytes that are similar to those found in disease. While some of them were expected (abnormalities in creatinine in patients with high genetic risk for chronic kidney disease), some other associations were novel (such as abnormalities in the metabolite 4-cholesten-3-one in patients with high genetic risk for PSC). Correlations were assessed by Glass' Δ (a measure of effect size evaluating the difference in standard deviations among groups); would ML add value in a study with this design? Unsupervised learning through association mining would be probably worth pursuing and may reveal other interesting findings arising from data.

A much promising multi-omics approach is that described by the European LifeTime Initiative, which aims to integrate single-cell sequencing techniques, imaging, and patient-derived experimental disease models by means of AI (100). Investigators behind the LifeTime Initiative have introduced the concept of *interceptive medicine*, which is the early interception of disease based on more accurate cellular and molecular diagnostics. In other words, the idea behind interceptive medicine is to combine several breakthrough technologies such as single-cell sequencing and DL to track the process of the disease in an unprecedented way, highlighting potential druggable pathways that are significant in the early phase of the disease before the fibrotic process occurs (for inflammatory disease) or the disease spreads throughout the body (for cancer). (Figure 3). Chronic inflammatory diseases (CIDs) like AiLDs are often detected late when tissues have already undergone extensive and non-reversible changes, hindering therapeutic options. This can be due to several reasons. We lack longitudinal tracking of cellular heterogeneity and molecular cell trajectories from a healthy to a diseased state, there is often fragmentation of approaches without systematic profiling of patients, computational algorithms for integration are still under development, and drugs are still given to most patients without a clear insight into the precise molecular abnormalities within the specific subject. Despite adopting a systematic approach and involving thousands of subjects that underwent deep phenotyping, the previously mentioned study from Wainberg and colleagues still suffers from some of the issues raised by LifeTime investigators. For example, longitudinal analyte data were collapsed to their median values, because genetic risk does not change over time; transcriptomics profiling was not included, and single-cell technologies were also not employed; finally, no ML or other AI algorithms were used (101).

The LifeTime initiative aims to develop cutting-edge pipelines that incorporate different single-cell technologies, to investigate cellular heterogeneity, and spatial molecular information, to better define the location of disease cells within the tissue. This information will be collected longitudinally by small/liquid biopsies and integrated with EHRs of patients. Organoid models derived from healthy and diseased individuals will represent the experimental arm of the pipeline. It is self-evident that the amount of information generated will need dedicated bioinformatics skills and infrastructures and, ultimately, ML. One of the envisioned final goals is the adoption of AI-based systems that will aid clinical decisions in everyday practice. This project identified five pillars: cancer, neurological disease, infections, cardiovascular diseases, and

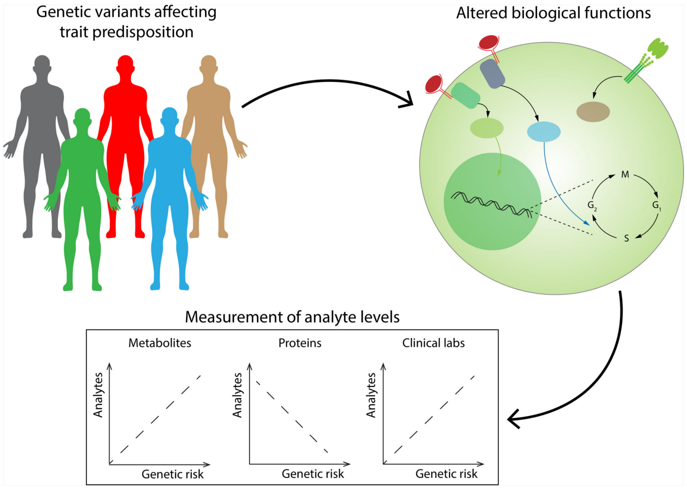
Genetic variants affecting trait predisposition
Altered biological functions
M
G2
G1
S
Measurement of analyte levels
Metabolites
Proteins
Clinical labs
Analytes
Genetic risk
Analytes
Genetic risk
Analytes
Genetic risk

FIGURE 2
Healthy individuals with high genetic risk scores for a specific trait have already detectable abnormalities in several blood analytes. Genetic risk scores generated from risk variants for several traits have been associated with levels of plasma analytes (standard blood analytes, and proteomic and metabolic measurements), revealing that a nonnegligible level of dysregulation of these analytes can already be found in healthy subjects with high genetic risk. These approaches could be potentially leveraged for early detection of diseases.

CIDs. As regards the latter ones, where AiLDs fall within, the urgent target is the investigation of cellular heterogeneity and how this influences disease course and differences in treatment response.

A systematic approach devoted to CIDs has been proposed by the SYSCID (*a systems medicine approach to chronic inflammatory diseases*) consortium (102). The consortium acknowledges that the field of immune-mediated diseases is lagging behind cancer and cardiovascular areas in its shift toward precision medicine, due to some hurdles. The first one relates to the problem of missing heritability, the concept that the successful GWAS have identified many variants associated with the risk of developing CIDs but each with little impact *per se* (76); many strategies are currently suggested to fill this gap, including adoption of ML on top of classical statistical genetics (86). Yet, missing heritability does not affect only CIDs but most of the complex traits (76). A second issue is related to the fragmentation of diagnostic and therapeutic pathways for CIDs despite being characterized by overlap in their molecular risk map; SYSCID researchers advocate for the development of dedicated centers for inflammation medicine, where different specialists take care of these diseases, similarly to what has already occurred for cancer. The third issue is represented by the discrepancy between the complexity of omics data and the need for simple scores in clinical practice; this gap is still large for CIDs. Longitudinal tracking of what happens in diseased tissues is probably unfeasible, due to inaccessibility, calling in action blood biomarkers; to this end, the study of Wainberg et al. represents a good proxy for future endeavors. The cultural paradigm shift needed for the birth of System Immunology as a discipline is to move from a hypothesis-driven approach studying single molecules, single-cell types, etc. toward a hypothesis-free integration of different layers of information: this is where ML may play a key role thanks to its characteristics. There is evidence that heterogeneity occurs at the interindividual level [e.g., the identification of cell-type-specific molecular quantitative trait loci (QTLs) that are dependent on different genetic variants] and at the intraindividual level (thanks to the characterization of different populations of cells within tissues by employing single-cell sequencing). Like the LifeTime Initiative, SYSCID also works under the European research scheme of Horizon2020 and focuses on three paradigmatic autoimmune diseases: rheumatoid arthritis, systemic lupus erythematosus, and inflammatory bowel disease. Five layers of data will be available for approximately 50,000 individuals: SNP variants, DNA methylome, transcriptome, immunoglobulin glycome, and gut microbiome. Canonical statistical modeling and novel ML techniques will be used to identify biomarkers, subtypes of disease, predictive models for tailored treatments, and novel reprogramming strategies.

Supervised learning could be used to separate groups according to clinical parameters of interests, such as treatment response; this will be likely a complementary or subsequent approach, since it requires a certain pre-test hypothesis. Unsupervised learning will be crucial in aggregating individuals based on the different layers of information. Kobak et al. have recently described pitfalls that may possibly occur in single-cell RNA sequencing data analysis (103); the addition of multiple layers of information will add even further complexity to data reduction algorithms. Multi-scale models will potentially enable to predict disease phenotypes at the cellular level (100).

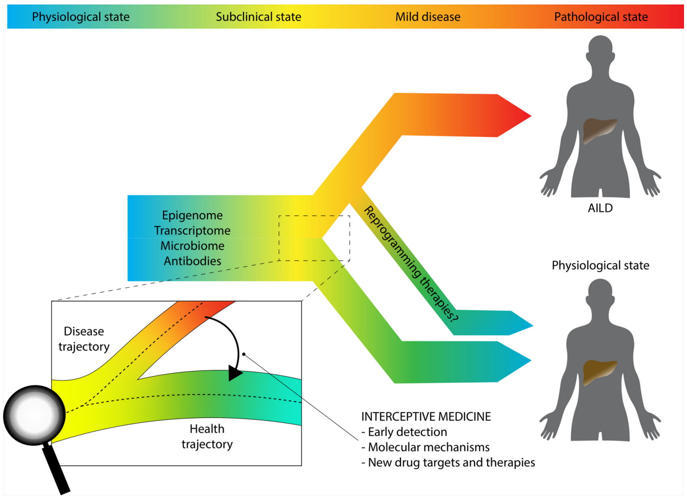

FIGURE 3
Interceptive medicine. The paradigm shift to study complex chronic immune-mediated traits proposed by the LifeTime Initiative and the SYSCID Consortium.

To summarize, making multi-omics integration a process that generates valuable scientific knowledge to translate in the clinic requires the collaboration among several scientists from different fields (4). Yet, the potential output of such an approach is potentially disruptive for the field of immune-mediated diseases and AiLDs more specifically.

3.6 Opening the black box: The importance of explainable AI

One of the fathers of the DL paradigm, Yoshua Bengio, recently highlighted, in a keynote speech at the IEEE World Congress on Computational Intelligence, how DL, for the time being, has been good in dealing with *a subset* of relevant activities. According to Bengio, the subset is related to the realm of *intuition*, rather than *explanation*: algorithms may perform well, but still struggle in *explaining why they perform well*.

Following Daniel Kahneman's classification of human intelligence (104), Bengio asserts that ML already achieved good results in emulating the behavior of what Kahneman refers to as *System 1* (which leads to intuitive unconscious decisions for human beings), while its path towards the emulation of *System 2* (that leads humans to deliberate and make conscious decisions) is still ongoing.

This is why many researchers focus on the development and on the extension of rule-based modeling techniques, so that each prediction is motivated by the rule(s) determining it. This is the case, for instance, of decision trees (105) and the logic learning machine (106). Considering the importance of keeping clinical experts at the core of the decision process in medicine, this kind of approach, focused on *interpretability*, may be of particular interest in the field.

DL has outstanding potential, but it is challenging to interpret DL systems and translate them in clinical practice. The lack of explainability makes it difficult to identify and tackle biases present in training sets. Most importantly, the research aim should be clear. If the goal is automation and technological aid to humans, lack of interpretability is probably much less important. Yet, when models are devised to make predictions that can change medical decisions, it is ethically difficult to accept a model that cannot be tracked in its work (107, 108). Furthermore, trustability is likely to be proportional to interpretability: it is less likely that a clinician would trust a model if he/she is not able to understand even a tiny part of it.

The encounter between ML and the biomedical field forces both specialists to learn something that is not typical of their domain (109). If clinicians will progressively be surrounded by concepts related to ML and its applications, ML scientists should work to fill some gaps that are still present in the ML literature (4). Many core statistical concepts, like calibration of estimates, confidence of estimates, or power calculation, have not been incorporated in ML models.

Overall, we acknowledge that these statements could be outpaced by changes in the field, which is moving rapidly and may pose new challenges, making these ones rapidly outdated.

4 Hurdles, limitations, and pitfalls

High expectations behind AI do exist (13); yet, we are progressively learning that several issues have to be tackled

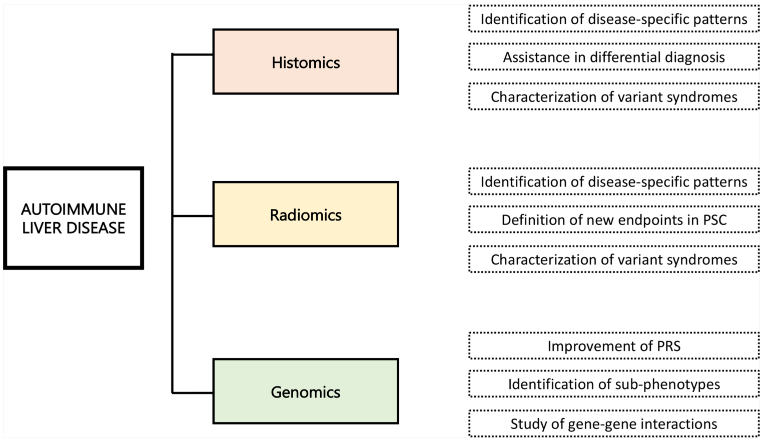
AUTOIMMUNE LIVER DISEASE
Histomics
Identification of disease-specific patterns
Assistance in differential diagnosis
Characterization of variant syndromes
Radiomics
Identification of disease-specific patterns
Definition of new endpoints in PSC
Characterization of variant syndromes
Genomics
Improvement of PRS
Identification of sub-phenotypes
Study of gene-gene interactions

FIGURE 4
Future applications of artificial intelligence in the field of autoimmune liver disease.

before full clinical implementation is possible (109). We can divide these issues into those concerning model development and those regarding model deployment (6, 110). It is also important to mention the difficulties related with the de-identification process required by data protection laws, especially relevant for the genetic field (85).

We mentioned the need for data standardization and explainability of models. It is also worth mentioning the issue of having a diverse dataset to have more generalized models; similarly to what is known in population genetics, where most of GWAS have been performed in populations of European origin, under-representing most of the other ancestries (111), there is a risk for training models in homogeneous populations where demographic factors are specific and, importantly, the etiology of the liver disease is different. To make an example, a model for non-invasive prediction of liver fibrosis trained in a US-derived cohort, where non-alcoholic fatty liver disease is the most prevalent cause, would probably fall short when applied in a Taiwanese cohort where chronic hepatitis B is leading. The need for diversity poses even a greater challenge for rare diseases like AiLDs, stressing the importance of creating big multicenter consortia and collaborative efforts to collect large-scale data.

Reproducibility of ML studies is a hot topic. There is evidence that the majority of studies do not follow a rigorous and consistent methodology (112, 113). Researchers in the field of computer science applied to medicine have been developing guidelines and standard methods that should be followed (114).

Yet, the field should also start focusing more on clinical deployment rather than doing only retrospective analysis for validation purposes (110). Digital transition will require building new infrastructures, training healthcare workforce, and involving patients in this process. Wherever possible, AI-based models should aim toward liberating the healthcare personnel from repetitive tasks to have more time to spend with their patients (13). Novel data sharing technologies are also required, to balance the need for data protection and to offer at the same time a large amount of data to train the algorithms. Blockchain-based swarm learning seems to offer a quite promising approach to this end (115, 116).

While AiLDs will share general hurdles for the implementation of AI in medicine, there are also disease-specific obstacles. We believe that AiLDs share with other rare conditions the difficulty in having sufficient sample sizes for these data-hungry new methodologies; despite the birth of worldwide consortia, the field will never reach the numbers of cardiovascular or cancer specialties. The wide heterogeneity of phenotypes seen in the clinic represent another peculiar obstacle for supervised approaches, which require clear-cut phenotypes as outputs. On the other hand, unsupervised learning techniques, although much awaited in AiLD for their capacity to detect patterns within data and pinpoint sub-phenotypes that are only intuitively noticed by clinicians, need bigger samples than supervised ones.

Overall, we are at a unique juncture in the history of medicine, with new technological avenues bringing together new challenges.

5 Conclusions

The application of ML to big data in medicine, and more specifically in AiLDs, is challenging. Figure 4 recapitulates the future applications of AI in the field of AILD. Major scientific, infrastructural, and cultural changes are needed (102). International endeavors should be implemented, to cut costs and address many, if not all, immune-mediated diseases altogether, breaking cultural barriers between clinicians and computational scientists, and educating the public about the benefit of the access to big personal health datasets rather than focusing on the privacy issues. It is essential to cut down the barriers to accessing genomic data derived from direct-to-consumer testing, such as 23andMe and other initiatives, which have been largely underutilized so far (117). If it is true that rare diseases suffer from minor availability of large amount of data, it would be of great value to take advantage of data generated by wearable devices and mobile phones, or from genetic analyses that individuals perform without clinical indication (e.g., 23andMe and others). Moreover, many data regarding rare conditions are already present in public repositories but may be scattered among different platforms and datasets.

Explainability of AI will be probably required by regulatory agencies to support systemic approaches and to accept their future translation into clinical protocols and pathways (118). Another caveat is that ML mostly finds correlation and is a hypothesis-generating tool; this means that it can open new doors, but all new ideas should still be tested experimentally, either in the laboratory or in clinical trials.

Nevertheless, despite all these potential obstacles to the application of AI in biomedicine, we should realize that this process could revolutionize the way we diagnose and treat our patients, which is the ultimate goal of translational research. Successful implementation will require investment in healthcare workforce education and technological infrastructures, together with involvement of patients and the creation of a culture of innovation and learning. To do so, we advocate that health services together with regulatory bodies should create a robust framework able to leverage the opportunities and address all the challenges that AI provides.

Author contributions

The first draft of the manuscript was prepared by AG with the support of MS, LC, and DV. CM has created and processed all figures. All authors contributed to manuscript revision, and read and approved the submitted version.

Funding

This research was partially supported by the Italian Ministry of University and Research (MIUR) - Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to clinic).

Acknowledgments

Alessio Gerussi, Miki Scaravaglio, Laura Cristoferi, Pietro Invernizzi, and Marco Carbone are members of the European Reference Network on Hepatological Diseases (ERN RARE-LIVER). This work has been supported by grants of the Italian Ministry of Health in the role of auto-reactive hepatic natural killer cells in the pathogenesis of primary biliary cholangitis (PE-2016-02363915) and in the biocompatible nano-assemblies to increase the safety and the efficacy of steroid treatment against liver inflammation (GR-2018-12367794). The authors thank AMAF Monza ONLUS and AIRCS for the unrestricted research funding.

Conflict of interest

Author DV is employed by Rulex.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Mieli-Vergani G, Vergani D, Czaja AJ, Manns MP, Krawitt EL, Vierling JM, et al. Autoimmune hepatitis. *Nat Rev Dis Prim* (2018) 4:18017. doi: 10.1038/nrdp.2018.17

2. Leung KK, Deeb M, Hirschfield GM. Review article: pathophysiology and management of primary biliary cholangitis. *Aliment Pharmacol Ther* (2020) 52:1150–64. doi: 10.1111/apt.16023

3. Karlsen TH, Folseraas T, Thorburn D, Vesterhus M. Primary sclerosing cholangitis – a comprehensive review. *J Hepatol* (2017) 67:1298–323. doi: 10.1016/j.jhep.2017.07.022

4. Littmann M, Selig K, Cohen-Lavi L, Frank Y, Hönigschmid P, Kataka E, et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat Mach Intell* (2020) 2:18–24. doi: 10.1038/s42256-019-0139-8

5. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev* (1959) 3(3):210–29. doi: 10.1147/rd.33.0210

6. Nam D, Chapiro J, Paradis V, Seraphin TP, Kather JN. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction. *JHEP Rep* (2022) 4:100443. doi: 10.1016/j.jhepr.2022.100443

7. Bzdok D, Altman N, Krzywinski M. Points of significance: Statistics versus machine learning. *Nat Methods* (2018) 15:233–4. doi: 10.1038/nmeth.4642

8. Kleppe A, Skrede OJ, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* (2021) 21:199–211. doi: 10.1038/s41568-020-00327-9

9. Rosenblatt F. *The perceptron: A perceiving and recognizing automaton*. Buffalo, New York: Cornell Aeronautical Laboratory (1957). Report 85-60-1.

10. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* (1995) 20:273–97. doi: 10.1007/BF00994018

11. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* (2015) 521:436–44. doi: 10.1038/nature14539

12. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge: MIT Press (2016).

13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7

14. Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* (2020) 72:2000–13. doi: 10.1002/hep.31207

15. Haas ME, Pirruccello JP, Friedman SN, Wang M, Emdin CA, Ajmera VH, et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genomics* (2021) 1:100066. doi: 10.1016/j.xgen.2021.100066

16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017) 542:115–8. doi: 10.1038/nature21056

17. Gui X, Bazarova A, Del Amor R, Vieth M, de Hertogh G, Villanacci V, et al. PICaSSO histologic remission index (PHRI) in ulcerative colitis: development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system. *Gut* (2022) 71:889–98. doi: 10.1136/gutjnl-2021-326376

18. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* (2019) 572:116–9. doi: 10.1038/s41586-019-1390-1

19. Gerussi A, Verda D, Bernasconi DP, Carbone M, Komori A, Abe M, et al. Machine learning in primary biliary cholangitis: a novel approach for risk stratification. *Liver Int* (2021), 42(3): 615–27. doi: 10.1111/liv.15141

20. Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA - J Am Med Assoc* (2019) 321(20):2003–17. doi: 10.1001/jama.2019.5791

21. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. *Nat Genet* (2019) 51:12–8. doi: 10.1038/s41588-018-0295-5

22. Melo RCN, Raas MWD, Palazzi C, Neves VH, Malta KK, Silva TP. Whole slide imaging and its applications to histopathological studies of liver disorders. *Front Med* (2020) 6:310. doi: 10.3389/fmed.2019.00310

23. Bozorgtabar B, Mahapatra D, Zlobec I, Rau TT, Thiran JP. Editorial: Computational pathology. *Front Med* (2020) 7:245. doi: 10.3389/fmed.2020.00245

24. Kather JN, Calderaro J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat Rev Gastroenterol Hepatol* (2020) 17:591–2. doi: 10.1038/s41575-020-0343-3

25. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer* (2021) 124:686–96. doi: 10.1038/s41416-020-01122-x

26. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med* (2021) 27:775–84. doi: 10.1038/s41591-021-01343-4

27. Saco A, Diaz A, Hernandez M, Martinez D, Montironi C, Castillo P, et al. Validation of whole-slide imaging in the primary diagnosis of liver biopsies in a university hospital. *Dig liver Dis Off J Ital Soc Gastroenterol Ital Assoc Study Liver* (2017) 49:1240–6. doi: 10.1016/j.dld.2017.07.002

28. 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) 162:1948–1961.e7. doi: 10.1053/j.gastro.2022.02.025

29. Jie-Yi S, Wang X, Ding G-Y, Zhou D, Han J, Guan Z, et al. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. *Gut* (2021) 70:951–61. doi: 10.1136/gutjnl-2020-320930

30. Marini N, Marchesin S, Otálora S, Wodzinski M, Caputo A, van Rijthoven M, et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *NPJ Digit Med* (2022) 5:102. doi: 10.1038/s41746-022-00635-4

31. Schömig-Markiefka B, Pryalukhin A, Hulla W, Bychkov A, Fukuoka J, Madabhushi A, et al. Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod Pathol* (2021) 34:2098–108. doi: 10.1038/s41379-021-00859-x

32. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun* (2021) 12:4423. doi: 10.1038/s41467-021-24698-1

33. Ren J, Hacihaliloglu I, Singer EA, Foran DJ, Qi X. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Front Bioeng Biotechnol* (2019) 7:102. doi: 10.3389/fbioe.2019.00102

34. Iglesias LL, Bellón PS, Del Barrio AP, Fernández-Miranda PM, González DR, Vega JA, et al. A primer on deep learning and convolutional neural networks for clinicians. *Insights Imaging* (2021) 12:117. doi: 10.1186/s13244-021-01052-z

35. Nakanuma Y, Zen Y, Harada K, Sasaki M, Nonomura A, Uehara T, et al. Application of a new histological staging and grading system for primary biliary cirrhosis to liver biopsy specimens: Interobserver agreement. *Pathol Int* (2010) 60:167–74. doi: 10.1111/j.1440-1827.2009.02500.x

36. de Vries EMG, de Krijger M, Färkkilä M, Arola J, Schirmacher P, Gotthardt D, et al. Validation of the prognostic value of histologic scoring systems in primary sclerosing cholangitis: An international cohort study. *Hepatology* (2017) 65:907–19. doi: 10.1002/hep.28963

37. de Boer YS, van Nieuwkerk CM, Witte BI, Mulder CJ, Bouma G, Bloemena E. Assessment of the histopathological key features in autoimmune hepatitis. *Histopathology* (2015) 66:351–62. doi: 10.1111/his.12558

38. Tiniakos DG, Brain JG, Bury YA. Role of histopathology in autoimmune hepatitis. *Dig Dis* (2015) 33(suppl 2):53–64. doi: 10.1159/000440747

39. Rahim MN, Miquel R, Heneghan MA. Approach to the patient with acute severe autoimmune hepatitis. *JHEP Rep* (2020) 2(6):100149. doi: 10.1016/j.jhepr.2020.100149

40. Björnsson E, Talwalkar J, Treeprasertsuk S, Kamath PS, Takahashi N, Sanderson S, et al. Drug-induced autoimmune hepatitis: Clinical characteristics and prognosis. *Hepatology* (2010) 51:2040–8. doi: 10.1002/hep.23588

41. Liu Y, Wang H, Taylor M, Cook C, Martínez-Berdeja A, North JP, et al. Classification of human chronic inflammatory skin disease based on single-cell immune profiling. *Sci Immunol* (2022) 7:eabl9165. doi: 10.1126/sciimmunol.abl9165

42. Gong C, Anders RA, Zhu Q, Taube JM, Green B, Cheng W, et al. Quantitative characterization of CD8+ T cell clustering and spatial heterogeneity in solid tumors. *Front Oncol* (2018) 8:649. doi: 10.3389/fonc.2018.00649

43. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* (2011) 3:108ra113. doi: 10.1126/scitranslmed.3002564

44. Kong J, Cooper LAD, Wang F, Gao J, Teodoro G, Scarpace L, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PloS One* (2013) 8:e81049–9. doi: 10.1371/journal.pone.0081049

45. European Association for the Study of the Liver. EASL clinical practice guidelines: The diagnosis and management of patients with primary biliary cholangitis. *J Hepatol* (2017) 145:167–72. doi: 10.1016/j.jhep.2017.03.022

46. De Vries EMG, Verheij J, Hubscher SG, Leeflang MMG, Boonstra K, Beuers U, et al. Applicability and prognostic value of histologic scoring systems in primary sclerosing cholangitis. *J Hepatol* (2015) 63:1212–9. doi: 10.1016/j.jhep.2015.06.008

47. Carbone M, Nardi A, Flack S, Carpino G, Varvaropoulou N, Gavrila C, et al. Pretreatment prediction of response to ursodeoxycholic acid in primary biliary cholangitis: development and validation of the UDCA response score. *Lancet Gastroenterol Hepatol* (2018) 1253:1–9. doi: 10.1016/S2468-1253(18)30163-8

48. Carpino G, Cardinale V, Folseraas T, Overi D, Floreani A, Franchitto A, et al. Hepatic Stem/Progenitor cell activation differs between primary sclerosing and primary biliary cholangitis. *Am J Pathol* (2018) 188:627–39. doi: 10.1016/j.ajpath.2017.11.010

49. Topol EJ. A decade of digital medicine innovation. *Sci Transl Med* (2019) 11:eaaw7610. doi: 10.1126/scitranslmed.aaw7610

50. Wei J, Jiang H, Gu D, Niu M, Fu F, Han Y, et al. Radiomics in liver diseases: Current progress and future opportunities. *Liver Int* (2020) 40:2050–63. doi: 10.1111/liv.14555

51. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (2017) 14:749–62. doi: 10.1038/nrclinonc.2017.141

52. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036

53. Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH. Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology* (2021) 73:2546–63. doi: 10.1002/hep.31603

54. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5

55. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, von Tengg-Kobligk H, et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiol Artif Intell* (2020) 2:e190043. doi: 10.1148/ryai.2020190043

56. Mulinacci G, Cristoferi L, Palermo A, Lucà M, Gerussi A, Invernizzi P, et al. Risk stratification in primary sclerosing cholangitis. *Minerva Gastroenterol Dietol* (2020). doi: 10.23736/S1121-421X.20.02821-4

57. Andres A, Montano-Loza A, Greiner R, Uhlich M, Jin P, Hoehn B, et al. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS One* (2018) 13:e0193523. doi: 10.1371/journal.pone.0193523

58. Eaton JE, Vesterhus M, McCauley BM, Atkinson EJ, Schlicht EM, Juran BD, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: A derivation and validation study using machine learning. *Hepatology* (2020) 71:214–24. doi: 10.1002/hep.30085

59. Corpechot C, Gaouar F, El Naggar A, Kemgang A, Wendum D, Poupon R, et al. Baseline values and changes in liver stiffness measured by transient elastography are associated with severity of fibrosis and outcomes of patients with primary sclerosing cholangitis. *Gastroenterology* (2014) 146:970–979.e6. doi: 10.1053/j.gastro.2013.12.030

60. Eaton JE, Dzyubak B, Venkatesh SK, Smyrk TC, Gores GJ, Ehman RL, et al. Performance of magnetic resonance elastography in primary sclerosing cholangitis. *J Gastroenterol Hepatol* (2016) 31:1184–90. doi: 10.1111/jgh.13263

61. Ni Mhuircheartaigh JM, Lee KS, Curry MP, Pedrosa I, Mortele KJ. Early peribiliary hyperenhancement on MRI in patients with primary sclerosing cholangitis: Significance and association with the Mayo risk score. *Abdom Radiol (New York)* (2017) 42:152–8. doi: 10.1007/s00261-016-0847-z

62. Ehlken H, Wroblewski R, Corpechot C, Arrivé L, Lezius S, Hartl J, et al. Spleen size for the prediction of clinical outcome in patients with primary sclerosing cholangitis. *Gut* (2016) 65:1230 LP–1232. doi: 10.1136/gutjnl-2016-311452

63. Goldfinger MH, Ridgway GR, Ferreira C, Langford CR, Cheng L, Kazimianec A, et al. Quantitative MRCP imaging: Accuracy, repeatability, reproducibility, and cohort-derived normative ranges. *J Magn Reson Imaging* (2020) 52:807–20. doi: 10.1002/jmri.27113

64. Janowski K, Shumbayawonda E, Cheng L, Langford C, Dennis A, Kelly M, et al. Quantitative multiparametric MRI as a non-invasive stratification tool in children and adolescents with autoimmune liver disease. *Sci Rep* (2021) 11:15261. doi: 10.1038/s41598-021-94754-9

65. Gilligan LA, Trout AT, Lam S, Singh R, Tkach JA, Serai SD, et al. Differentiating pediatric autoimmune liver diseases by quantitative magnetic resonance cholangiopancreatography. *Abdom Radiol (New York)* (2020) 45:168–76. doi: 10.1007/s00261-019-02184-z

66. Ismail MF, Hirschfield GM, Hansen B, Tafur M, Elbanna KY, Goldfinger MH, et al. Evaluation of quantitative MRCP (MRCP+) for risk stratification of primary sclerosing cholangitis: comparison with morphological MRCP, MR elastography, and biochemical risk scores. *Eur Radiol* (2021) 32(1):67–77. doi: 10.1007/s00330-021-08142-y

67. Selvaraj EA, Ba-Ssalamah A, Poetter-Lang S, Ridgway GR, Brady JM, Collier J, et al. A quantitative magnetic resonance cholangiopancreatography metric of intrahepatic biliary dilatation severity detects high-risk primary sclerosing cholangitis. *Hepatol Commun* (2021) 6(4):795–808. doi: 10.1002/hep4.1860

68. Arndtz K, Shumbayawonda E, Hodson J, Eddowes PJ, Dennis A, Thomaides-Brears H, et al. Multiparametric magnetic resonance imaging, autoimmune hepatitis, and prediction of disease activity. *Hepatol Commun* (2021) 5(6):1009–20. doi: 10.1002/hep4.1687

69. Janowski K, Shumbayawonda E, Dennis A, Kelly M, Bachtiar V, DeBrota D, et al. Multiparametric MRI as a noninvasive monitoring tool for children with autoimmune hepatitis. *J Pediatr Gastroenterol Nutr* (2021) 72:108–14. doi: 10.1097/MPG.0000000000002930

70. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat Rev Genet* (2018) 19:110–24. doi: 10.1038/nrg.2017.101

71. Engel B, Laschtowitz A, Janik MK, Junge N, Baumann U, Milkiewicz P, et al. Genetic aspects of adult and pediatric autoimmune hepatitis: a concise review. *Eur J Med Genet* (2021) 64(6):104214. doi: 10.1016/j.ejmg.2021.104214

72. Gerussi A, Asselta R, Invernizzi P. Genetics of primary biliary cholangitis. *Clin Liver Dis* (2022) 26(4):571–82. doi: 10.1016/j.cld.2022.06.002

73. Jiang X, Karlsen TH. Genetics of primary sclerosing cholangitis and pathophysiological implications. *Nat Rev Gastroenterol Hepatol* (2017) 14:279–95. doi: 10.1038/nrgastro.2016.154

74. Gerussi A, Carbone M, Corpechot C, Schramm C. The genetic architecture of primary biliary cholangitis. *Eur J Med Genet* (2021) 64:104292. doi: 10.1016/j.ejmg.2021.104292

75. Papatheodoridis GV, Lekakis V, Voulgaris T, Lampertico P, Berg T, Chan HLY, et al. Hepatitis b virus reactivation associated with new classes of immunosuppressants and immunomodulators: A systematic review, meta-analysis, and expert opinion. *J Hepatol* (2022) S0168-8278(22)02935-X. doi: 10.1016/j.jhep.2022.07.003

76. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature* (2009) 461:747–53. doi: 10.1038/nature08494

77. Ellinghaus D. How genetic risk contributes to autoimmune liver disease. *Semin Immunopathol* (2022) 44:397–410. doi: 10.1007/s00281-022-00950-8

78. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* (2001) 69:124–37. doi: 10.1086/321272

79. Wainschtein P, Jain D, Zheng Z, Cupples LA, Shadyab AH, McKnight B, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* (2022) 54:263–73. doi: 10.1038/s41588-021-00997-7

80. Jiang X, Bergquist A, Löscher B-S, Venkatesh G, Mold JE, Holm K, et al. A heterozygous germline CD100 mutation in a family with primary sclerosing cholangitis. *Sci Transl Med* (2021) 13:eabb0036. doi: 10.1126/scitranslmed.abb0036

81. Haisma S-M, Weersma RK, Joosse ME, de Koning BAE, de Meij T, Koot BGP, et al. Exome sequencing in patient-parent trios suggests new candidate genes for early-onset primary sclerosing cholangitis. *Liver Int Off J Int Assoc Study Liver* (2021) 41:1044–57. doi: 10.1111/liv.14831

82. Schrider DR, Kern AD. Supervised machine learning for population genetics: A new paradigm. *Trends Genet* (2018) 34:301–12. doi: 10.1016/j.tig.2017.12.005

83. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. *Front Genet* (2019) 10:267. doi: 10.3389/fgene.2019.00267

84. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PloS Genet* (2014) 10:e1004754. doi: 10.1371/journal.pgen.1004754

85. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. *Nat Rev Genet* (2020) 21:615–29. doi: 10.1038/s41576-020-0257-5

86. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* (2019) 20:467–84. doi: 10.1038/s41576-019-0127-1

87. Phillips PC. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* (2008) 9:855–67. doi: 10.1038/nrg2452

88. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* (2009) 10:392–404. doi: 10.1038/nrg2579

89. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. *Front Genet* (2015) 6:285. doi: 10.3389/fgene.2015.00285

90. Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinf* (2009) 10 Suppl 1:S65–5. doi: 10.1186/1471-2105-10-S1-S65

91. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* (2013) 14:507–15. doi: 10.1038/nrg3457

92. Tana MM-S, Klepper A, Lyden A, Pisco AO, Phelps M, McGee B, et al. Transcriptomic profiling of blood from autoimmune hepatitis patients reveals potential mechanisms with implications for management. *PloS One* (2022) 17: e0264307. doi: 10.1371/journal.pone.0264307

93. Manfredo Vieira S, Hiltensperger M, Kumar V, Zegarra-Ruiz D, Dehner C, Khan N, et al. Translocation of a gut pathobiont drives autoimmunity in mice and humans. *Science (80-)* (2018) 359:1156–61. doi: 10.1126/science.aar7201

94. Clemente JC, Manasson J, Scher JU. The role of the gut microbiome in systemic inflammatory disease. *BMJ* (2018) 360:j5145–5. doi: 10.1136/bmj.j5145

95. Ramachandran P, Matchett KP, Dobie R, Wilson-Kanamori JR, Henderson NC. Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. *Nat Rev Gastroenterol Hepatol* (2020) 17(8):457–72. doi: 10.1038/s41575-020-0304-x

96. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet* (2018) 19(5):299–310. doi: 10.1038/nrg.2018.4

97. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z

98. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data* (2019) 6:251. doi: 10.1038/s41597-019-0258-4

99. Wilkinson MD, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi: 10.1038/sdata.2016.18

100. Rajewsky N, Almouzni G, Gorski SA, Aerts S, Amit I, Bertero MG, et al. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* (2020) 587(7834):377–86. doi: 10.1038/s41586-020-2715-9

101. Wainberg M, Magis AT, Earls JC, Lovejoy JC, Sinnott-Armstrong N, Omenn GS, et al. Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. *Proc Natl Acad Sci* (2020) 117:21813 LP – 21820. doi: 10.1073/pnas.2001429117

102. Schultze JL, Rosenstiel P. Systems medicine in chronic inflammatory diseases. *Immunity* (2018) 48:608–13. doi: 10.1016/j.immuni.2018.03.022

103. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* (2019) 10:5416. doi: 10.1038/s41467-019-13056-x

104. Kahneman D. *Thinking, fast and slow*. Farrar, Straus and Giroux (2011).

105. Quinlan JR. Induction of decision tress. *Mach Learn* (1986) 1.1:81–106. doi: 10.1007/BF00116251

106. Verda D, Parodi S, Ferrari E, Muselli M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinf* (2019) 20:1–13. doi: 10.1186/s12859-019-2953-8

107. Kundu S. AI In medicine must be explainable. *Nat Med* (2021) 27:1328. doi: 10.1038/s41591-021-01461-z

108. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-)* (2019) 366:447 LP – 453. doi: 10.1126/science.aax2342

109. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0

110. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI In health and medicine. *Nat Med* (2022) 28(1):31–8. doi: 10.1038/s41591-021-01614-0

111. Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, et al. Getting genetic ancestry right for science and society. *Science* (2022) 376 (6590):250–2. doi: 10.1126/science.abm7530

112. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: A systematic review. *PloS Med* (2012) 9:e1001221. doi: 10.1371/journal.pmed.1001221

113. Vandewiele G, Dehaene I, Kovács G, Sterckx L, Janssens O, Ongenae F, et al. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. *Artif Intell Med* (2021) 111:101987. doi: 10.1016/j.artmed.2020.101987

114. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int J Med Inform* (2021) 153:104510. doi: 10.1016/j.ijmedinf.2021.104510

115. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* (2021) 594:265–70. doi: 10.1038/s41586-021-03583-3

116. Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med* (2022) 28(6):1232–9. doi: 10.1038/s41591-022-01768-5

117. Green ED, Gunter C, Biesecker LG, Di Francesco V, Easter CL, Feingold EA, et al. Strategic vision for improving human health at the forefront of genomics. *Nature* (2020) 586:683–92. doi: 10.1038/s41586-020-2817-4

118. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI–explainable artificial intelligence. *Sci Robot* (2019) 4:eaay7120. doi: 10.1126/scirobotics.aay7120