Title
Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity

Journal
Gut, 71(9)

ISSN
0017-5749

Authors
Uche-Anya, Eugenia
Anyane-Yeboa, Adjoa
Berzin, Tyler M
et al.

Publication Date
2022-09-01

DOI
10.1136/gutjnl-2021-326271

Peer reviewed

Artificial intelligence in gastroenterology and hepatology: how to advance clinical practice while ensuring health equity

Eugenia Uche-Anya ,[1] Adjoa Anyane-Yeboa,[1] Tyler M Berzin ,[2] Marzyeh Ghassemi,[3] Folasade P May[4]

[1]Division of Gastroenterology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[2]Center for Advanced Endoscopy, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, USA
[3]Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]Vatche and Tamar Manoukian Division of Digestive Diseases, UCLA Kaiser Permanente Center for Health Equity and Jonsson Comprehensive Cancer Center, University of California Los Angeles, Los Angeles, California, USA

Correspondence to
Dr Folasade P May, Vatche and Tamar Manoukian Division of Digestive Diseases, UCLA Kaiser Permanente Center for Health Equity and Jonsson Comprehensive Cancer Center, University of California Los Angeles, Los Angeles, CA 90095-6900, USA; fmay@mednet.ucla.edu

EU-A and AA-Y are joint first authors.

Received 7 October 2021
Accepted 19 April 2022

Check for updates

To cite: Uche-Anya E, Anyane-Yeboa A, Berzin TM, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2021-326271

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) systems are increasingly used in medicine to improve clinical decision-making and healthcare delivery. In gastroenterology and hepatology, studies have explored a myriad of opportunities for AI/ML applications which are already making the transition to bedside. Despite these advances, there is a risk that biases and health inequities can be introduced or exacerbated by these technologies. If unrecognised, these technologies could generate or worsen systematic racial, ethnic and sex disparities when deployed on a large scale. There are several mechanisms through which AI/ML could contribute to health inequities in gastroenterology and hepatology, including diagnosis of oesophageal cancer, management of inflammatory bowel disease (IBD), liver transplantation, colorectal cancer screening and many others. This review adapts a framework for ethical AI/ML development and application to gastroenterology and hepatology such that clinical practice is advanced while minimising bias and optimising health equity.

Key messages

- ⇒ Artificial intelligence (AI) and machine learning (ML) systems are increasingly used in medicine to improve clinical decision-making and healthcare delivery.
- ⇒ In gastroenterology and hepatology, studies have explored a myriad of opportunities for AI/ML applications which are already making the transition to bedside.
- ⇒ Despite these advances, there is a risk that biases and health inequities can be introduced or exacerbated by these technologies. If unrecognised, these technologies could generate or worsen systematic racial, ethnic and sex disparities when deployed on a large scale.
- ⇒ There are several mechanisms through which AI/ML could contribute to health inequities in gastroenterology and hepatology, including diagnosis of oesophageal cancer, management of inflammatory bowel disease (IBD), liver transplantation, colorectal cancer screening and many others.
- ⇒ This review adapts a framework for ethical AI/ML development and application to gastroenterology and hepatology such that clinical practice is advanced while minimising bias and optimising health equity.

INTRODUCTION: ARTIFICIAL INTELLIGENCE AND HEALTH EQUITY

Artificial intelligence (AI) and machine learning (ML) technologies can leverage massive amounts of data for predictive modelling in a wide variety of fields and are increasingly used to inform complex decision-making and clinical processes in healthcare.[1] Examples include computer vision-assisted mammograms to improve breast cancer detection,[2] models that predict respiratory decompensation in patients with COVID-19[3] and AI tools which predict length of stay, facilitate resource allocation and lower healthcare costs.[4]

In gastroenterology and hepatology, opportunities for AI/ML implementation are burgeoning. Recent studies have explored AI applications such as computer-aided detection (CADe) for diagnosis of premalignant and malignant GI lesions, prediction of treatment response in patients with inflammatory bowel disease (IBD), histopathological analysis of biopsy specimens, assessment of liver fibrosis severity in chronic liver disease, models for liver transplant allocation and others.[5–8]

AI-based systems are increasingly making the transition from research to bedside and have the potential to revolutionise patient care. However, these advances must be matched by corresponding regulatory and ethical frameworks developed by the Food and Drug Administration (FDA) and other agencies that oversee the intended and unintended consequences of their use.[9] Concerns have already been raised regarding the biases and health inequities that can be introduced or amplified when applying computer algorithms in healthcare.[10] For instance, a commercial algorithm applied to approximately 200 million patients in the USA was racially biased—white patients were preferentially enrolled in ‘high-risk care management programmes’ compared with black patients with similar risk scores, resulting in fewer healthcare dollars spent on black patients.[11] Another study demonstrated that an ML algorithm that predicts intensive care unit mortality and 30-day psychiatric readmission rates had poorer predictive performance for women and patients with public insurance.[12]

Prior to deploying and scaling AI/ML tools, it is critical to ensure that the risk of bias is minimised

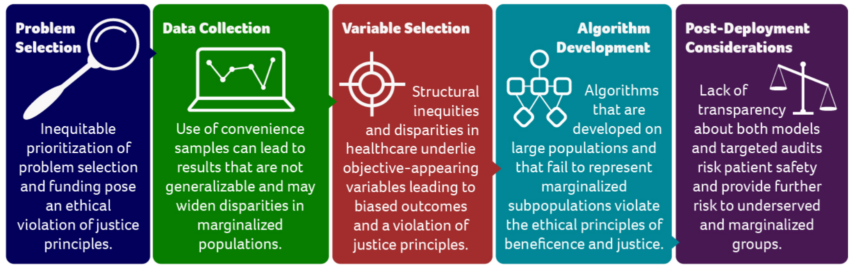

Figure 1 Mechanisms through which AI contributes to health inequities. Adapted from Chen *et al*.[14] AI, artificial intelligence.

and opportunities to promote health equity are amplified. In this work, we identify areas in gastroenterology and hepatology where algorithms could exacerbate disparities, and offer potential areas of opportunity for advancing health equity through AI/ML. For the purposes of this paper, health equity is centred on distributive justice to eliminate systematic racial, ethnic and sex disparities.[13]

FIVE KEY MECHANISMS THROUGH WHICH AI/ML CAN CONTRIBUTE TO HEALTH INEQUITIES

Early efforts to promote responsible and ethical applications of AI and ML in clinical medicine have revealed several mechanisms through which algorithms can introduce bias and exacerbate health inequities (figure 1).[10 14] It is essential for researchers and clinicians in gastroenterology and hepatology to understand these mechanisms as new technologies are being applied to our field.

The first theme is *disparities in clinical or research problem selection*[14]—research questions often target concerns in majority populations due to unequal funding availability and/or interest in the problem by industry, researchers, funders and grant review committees. As a result, we see critical racial, ethnic and sex disparities in the research problems that are prioritised and funded in AI/ML.[14 15]

The second theme is *bias in data collection*,[14] where collected data may capture a disproportionate share of one population group over another. This can result in algorithms that are not widely generalisable,[14 16–19] especially for individuals from traditionally under-represented and marginalised groups that are not commonly or appreciably represented in research databases.

The third theme is *bias due to variable selection*.[14] Variables and outcome measures may appear unbiased on initial evaluation even though they are proxies for, or confounded by, explicit or implicit biases against under-represented or marginalised groups.

The fourth theme is *bias in algorithm development*.[14] In this case, the assumptions made and used by the research team lead to inherently biased models or models that are overfitted to narrow training data.[13 15 17] Additionally, the performance metrics of AI/ML model training, such as area under the curve, are not inherently optimal for equitable performance in diverse populations.[14 20]

Finally, *inequities may result from post-deployment considerations*.[14] Even a potentially unbiased AI tool may lead to biased behaviour when deployed in the clinical setting. It is important to consider (1) how a tool will perform in a disease that has different conditional distributions in a population and (2) the potentially negative human–computer interactions that may occur. For instance, providers may follow an AI/ML-generated treatment recommendation when it confirms their biased beliefs but disregard treatment recommendations that do not conform to their beliefs.[14 21] The impacts of the algorithm should be evaluated in population subgroups to assess differences in clinical behaviour, performance and outcomes by sociodemographic factors, rather than at the population level alone.[14]

AI IN GASTROENTEROLOGY AND HEPATOLOGY: IMPACTS ON HEALTH EQUITY

We have chosen specific clinical examples from the literature to illustrate the specific ways existing AI/ML algorithms may already exacerbate bias and inequities within the fields of gastroenterology and hepatology (table 1).

Oesophageal cancer

Prevention and early recognition of oesophageal cancer is an area in which AI may hold particular promise, and there has been meaningful research progress in this area. In the USA, the vast majority of oesophageal cancer research focuses on technologies (with and without AI) to improve the early identification and treatment of Barrett's oesophagus and oesophageal adenocarcinoma (OAC).[22–26] Unfortunately, this emphasis on OAC in the USA primarily benefits white populations, who have the highest incidence and mortality from OAC.[27]

Oesophageal squamous cell carcinoma (OSCC) is more common than OAC and has a higher incidence and mortality in non-white populations in the USA and worldwide.[28 29] Specifically, black individuals in the USA have the highest incidence of OSCC at 4.9 per 100 000 people, followed by Asians at 1.9 per 100 000, and white individuals with the lowest rates at 1.4 per 100 000. These rates are higher than overall OAC rates and OAC rates among non-white individuals in the USA: 2.3 per 100 000 for white individuals, compared with 0.5 per 100 000 in black and Asian individuals.[28] Furthermore, AI research in OSCC has largely been performed in Asian countries, and thus, it is uncertain whether findings may be generalisable to black individuals or other population subgroups globally.[30–34]

This research disparity may be due to a *clinical or research problem selection bias* (theme 1); both researchers and funders should work to ensure more equity in problem selection. AI-based tools for early recognition of oesophageal cancer and precursor lesions have the potential to save many lives, but in the current state will largely impact white patients and not individuals from under-represented groups. It is imperative that we consider inclusive AI/ML research questions to avoid preferential development of technologies that consistently benefit one group over others.

Table 1 Types of bias observed in artificial intelligence (AI) in clinical medicine

Theme	Definition	Examples
Problem selection	Differential research priority and funding for issues that affect marginalised groups.	► AI has been used extensively to detect Barrett's oesophagus and oesophageal adenocarcinoma, which mainly affects white individuals. ► In contrast, AI applications in oesophageal squamous carcinoma—which is more prevalent in underserved populations—are under-researched.
Data collection	Inadequate representation of underserved groups in training datasets results in biased algorithms that yield inaccurate outputs for these subgroups.	► A model trained on Veteran's Health Administration electronic database to predict IBD flares may generate incorrect predictions for non-white populations who are under-represented in the training dataset.
Variable selection	Seemingly objective predictor and outcome variables that are included in a model may be confounded by or proxies for factors that lead to biased results.	► MELD exception points may appropriately prioritise patients with HCC on the transplant waitlist; however, this is confounded by the increased prevalence of HCC in men which leads to lower transplant rates for women. ► The inclusion of serum creatinine in the MELD leads to lower scores and transplant priority for women as serum creatinine underestimates renal dysfunction in women.
Algorithm development	Models are developed to recreate patterns in the training dataset and may not account for systemic biases.	► Racial and ethnic minorities are less likely to be referred for liver transplant and more likely to be offered a lower quality allograft. Predictive models could learn these patterns and propagate existing disparities.
Post-deployment considerations	Potentially unbiased AI tools may lead to biased outcomes when deployed in real life either due to differential conditional distribution of outcomes of interests across subpopulations.	► Computer vision has been shown to aid detection of traditional adenomas; however, there are limited data on proximal and sessile serrated lesions which are more prevalent in black individuals.

HCC, hepatocellular carcinoma; MELD, Model for End-Stage Liver Disease.

Inflammatory bowel disease

In IBD, AI/ML computer vision tools have been developed for endoscopic assessment of disease severity, to distinguish colitis from neoplasia, and to differentiate sporadic adenomas from non-neoplastic lesions.[6 35] AI algorithms have also been trained to predict treatment response and assess risk of disease recurrence.[35–37] AI has the potential to play an important role in IBD treatment decisions by predicting response earlier in the treatment course and guiding personalised therapy choices.

However, many AI/ML models developed in IBD have been created in largely white populations. For example, one study used AI to predict future corticosteroid use and hospitalisation in patients with IBD from a cohort of 20 368 patients at the Veterans Health Administration (VA).[36] The authors concluded that their model had the potential to predict IBD flares, improve patient outcomes and reduce healthcare costs. They also noted that their algorithm would be easy to implement at the point of care to individualise and tailor therapies for individual patients. The population that was used to derive the model was 93% male which may make predictions for female patients with IBD less relevant. The algorithm also included race as a predictor, though the dataset was racially skewed: the study population was 70% white, 8% black, 1.7% other and 19% unknown. This study was replicated in a large insurance-based cohort of 95 878 patients—though women were more adequately represented (57.1%), the patient population was still predominantly white (87.7%).[38]

While IBD was previously thought of as a disease that predominantly affects white individuals, there is now an increasing incidence in other racial and ethnic groups in the USA and worldwide.[39–41] In addition, IBD management and outcomes are worse for black and Latino patients compared with white patients, which should prompt increased research and clinical decision support for these groups.[42 43] In the VA study, the proportion of the population that was Hispanic/Latinx or South and East Asian was not included, despite the fact that these groups comprise an increasingly large share of the populations with IBD. While this study may be beneficial to the patient population served by the VA, it suggests that even in very large cohorts, there may be entrenched patterns of bias in *data collection* (theme 2): algorithms that do not include the rich diversity of patients with IBD can result in biased systems, care and outcomes, particularly if extrapolated to the general population.

Liver transplantation

There are numerous opportunities for AI/ML applications in hepatology, including the assessment of hepatic fibrosis progression, detection of non-alcoholic fatty liver disease, identification of patients at risk of hepatocellular carcinoma (HCC) and optimisation of organ transplant protocols.[6 44] As we explore the complex and opaque nature of emerging AI clinical prediction tools, it is important to recognise that bias can be encoded even in conventional prediction models, including simple, rule-based algorithms.

Prior to the adoption of the Model for End-Stage Liver Disease (MELD) score in 2002, the liver allocation process was fraught with variability, subjectivity and opportunities for manipulation, which resulted in inequities.[45] To address these shortcomings and standardise the organ allocation process, the United Network for Organ Sharing turned to the MELD—an algorithmic model which predicts 3-month survival rates in patients with cirrhosis—as a way to more fairly prioritise patients for liver transplantation.[46] Variables included in the model appear to be objective laboratory values—bilirubin, creatinine, international normalised ratio and sodium. Creatinine however underestimates renal dysfunction in women, leading to lower MELD scores compared with men with similar disease severity. This underestimation negatively impacts equitable organ allocation for liver transplant.[47–49]

A similar example occurs with MELD exception points—a system where patients with certain conditions that confer excess risk beyond that captured by the laboratory variables that comprise MELD (such as HCC) may accumulate points and advance their position on the transplant waitlist.[50] Review of data from Organ Procurement and Transplantation Network (OPTN) registries shows that at similar listing priority, patients with MELD exception points are less likely to die on the waitlist, more likely to receive a transplant and less likely to be women.[49 51 52] Part of this discrepancy is because HCC—the indication for MELD exception points in approximately 70%

of patients[49 52]—is two to four times more common in men.[53 54] Therefore, the inclusion of this variable in the model inadvertently deprioritises women and perpetuates sex disparities in liver transplantation: women are up to 20% less likely to receive a liver transplant and 8.6% more likely to die on the transplant waitlist.[52 55]

While conventional prediction models use few variables that appear to be transparent, high-capacity ML algorithms may employ innumerable variables from large volumes of data and identify highly complex non-linear patterns that are less comprehensible—that is, black box models—with the promise of increased predictive accuracy.[44 56] Regardless of the type of model used, the *variables selected* (theme 3) as inputs for these models—conventional and AI-based alike—may appear objective at face value but can unwittingly introduce bias and lead to inequitable outcomes as illustrated with the MELD.

Bias due to variable selection is intrinsically related to the introduction of bias during algorithm development. Datasets used for predictive modelling may have unintended encoded biases, which has the potential to generate biased algorithms in the *algorithm development phase* (theme 4) as ML models aim to fit the datasets on which they train. For example, review of OPTN registry data reveals that medically underserved groups are less likely to be referred for liver transplant, less likely to undergo liver transplantation and more likely to receive lower quality allografts compared with white patients.[57] Predictive models trained on such datasets could recreate these biases and amplify existing racial and ethnic liver transplantation disparities when deployed. Assigning transplant priorities based on predicted outcomes from biased models has huge ramifications for health equity in organ allocation. Identifying and rectifying biases after the model has been deployed can prove to be difficult—it took several years to show that an estimated glomerular filtration rate equation widely used to assign renal transplant priorities was biased against black patients.[58] Therefore, it is imperative that fairness and potential biases are addressed upfront.

Colorectal cancer screening

Colorectal cancer (CRC) prevention and control are major public health contributions by gastroenterologists. Effective CRC screening depends significantly on the endoscopist's ability to identify and remove high-risk colon and rectal polyps during colonoscopy. Adenoma detection rate (ADR) is a validated measure of colonoscopy quality and significant predictor of interval CRC risk.[59] Wide variability in ADR has been observed among endoscopists[60]: this contributes to suboptimal colonoscopy efficacy in preventing CRC incidence and deaths. Advances in ML have led to the application of computer vision to aid polyp detection during colonoscopy with data supporting the use of CADe to increase ADR.[61 62] Recently, the US FDA approved the first AI software based on ML to assist clinicians in the detection of colorectal polyps.[9] However, when implementing these tools, it is important to consider the *conditional distributions* (theme 5) of colorectal polyps across subpopulations.

Both proximal (right-sided) and sessile serrated lesions (sessile serrated polyps and serrated adenomas) are more challenging to detect as they can be flat and subtle compared with traditional adenomas.[63] Data are limited on the sensitivity of CADe for proximal and sessile serrated lesions; one study suggests lower sensitivity for sessile serrated lesions.[64] As black patients are more likely to have proximal polyps[65–68] and to have sessile serrated lesions,[69 70] CADe models trained primarily on traditional adenomas may have higher miss rates for precancerous lesions and be less effective for black patients. As black individuals have 20% higher CRC incidence, 40% higher CRC deaths and 30% higher interval CRC risk, CADe has the potential to exacerbate existing racial disparities if their ability to detect high-risk polyps is reduced among black individuals or other patient populations.[71 72]

It is essential to determine whether these AI/ML-powered models are also adequately trained to detect the high-risk polyps that are more commonly seen in black populations and in other populations who also suffer disproportionately from CRC. Optimising and validating these models and their miss rates across multiple and diverse populations have the potential to reduce variability in colonoscopy quality and improve racial disparities in CRC, especially as these technologies begin to gain FDA approval and are applied to diverse community settings. However, it is important to highlight that medically underserved and vulnerable populations often face barriers to accessing these clinically indicated tests to begin with. For instance, black patients are less likely to receive colonoscopy screening/surveillance,[73–76] surveillance imaging for HCC[77 78] and cross-sectional abdominal staging scans for pancreatic cancer.[79] This challenge further limits the opportunities for AI research to optimise these tests for diverse populations and promote health equity.

INCREASING EQUITY IN AI: POTENTIAL SOLUTIONS

It is imperative that we identify and implement pragmatic solutions that emphasise and optimise health equity in AI/ML development and application in gastroenterology and hepatology. Tools are needed to debias data collection, model training, model outputs and clinical application. The recently increased focus on equity in healthcare has motivated discussion about how to achieve these goals; these approaches are also urgently relevant to our field.[11 14 20 80] Potential solutions to the equity challenges we have highlighted in this piece include incorporating a health equity lens early and often in AI/ML research and development, increasing the diversity of patients involved in AI/ML clinical trials, regulatory standards for reporting, and pre-deployment and post-deployment auditing (table 2).

First, a health equity approach to AI/ML requires technically diverse research teams that are aware of how bias can creep into all aspects of the research continuum. Beyond this, gastroenterology and hepatology research teams that employ AI/ML methods should engage health equity experts early in their work so that potential sources of bias are identified early and are addressed in a robust and effective manner.

Second, it is vitally important to increase the diversity of patient populations who are involved in algorithm development and validation in gastroenterology and hepatology. Data collection in AI in our field is currently limited by overfitting and spectrum bias. Overfitting occurs when models are closely tailored to a training set, which can reduce overall generalisability of the model when other datasets are used.[81] Spectrum bias occurs when the datasets used to develop models do not reflect the diversity of the population they are meant to serve.[81 82] Datasets used for AI in gastroenterology and hepatology are often collected via retrospective or case–control design which poses risk for spectrum bias.[81 82] Ideally, all algorithms should be developed and tested using a population that reflects the racial, ethnic, age, sex and gender diversity of our society to maximise generalisability in routine practice. Historically, research studies have not been conducted in settings that regularly serve these populations resulting in their ongoing exclusion. Therefore, it is critical to consider where marginalised populations are being

Table 2 Approaches to eliminate bias in AI/ML

Appropriate research expertise	Involve health equity experts in the conception, development and deployment of AI/ML
Diverse study populations	Diversify study populations to adequately represent marginalised populations in training datasets. Convenience samples, such as datasets from electronic health records, claims data and so on, may not be adequately representative of marginalised groups.
Diverse study settings	Expand research locations to non-conventional settings where traditionally under-represented and vulnerable populations can be easily reached such as community health centres, faith-based organisations, barbershops, community service organisations and other settings.
Regulatory measures	Determine fair, clear, specific and quantifiable regulatory measures of inequitable outcomes. Researchers should be required to report descriptive data on study populations by sex, race, ethnicity as long as privacy is protected. Standards should be consistent across regulatory bodies, peer-reviewed scientific journals and gastroenterology/hepatology professional societies.
Pre-deployment auditing	Mandate auditing processes and sensitivity analyses to assess algorithmic performance across subpopulations in the pre-deployment phases.
Post-deployment auditing	Establish auditing processes to assess algorithmic performance across subpopulations in the post-deployment phase and pathways for rapidly mitigating bias if discovered in the post-deployment phase.

AI, artificial intelligence; ML, machine learning.

served and how best to reach them, both in AI and non-AI contexts. Partnerships between gastroenterology practices, clinics and health centres that provide care for these populations can be leveraged to extend reach and promote generalisability when conducting research studies to advance equity. In addition, non-traditional settings where vulnerable populations routinely receive services should be considered to diversify representation in AI/ML studies and ensure equity in algorithm performance. Furthermore, models should be externally validated with new patient populations and datasets to limit the potential for spectrum bias and overfitting.[81]

Beyond diversifying the training data, it is crucial that labels (or data classifiers) used in prediction models are adequately representative of the desired outcome alone and are independent of societal inequities.[11] It is also important to carefully consider the different conditional distributions of labels across subgroups and any variations in how they are classified and measured—these may suggest a need for optimising benchmarks or developing separate models for different subgroups.[83] Some tools like Datasheets for Datasets[84] or Model Cards for Model Reporting[85] do exist, but identifying the precise cause of bias can be challenging and requires careful audits by multidisciplinary teams.

A third focus should be on regulatory standards. Mandating explicit reporting of descriptive data of the patient populations used in AI/ML development—such as race, ethnicity, income, insurance and sex—is a necessary step, as long as privacy is protected. Doing so enables a clear assessment of appropriate representation in the algorithm's training dataset and the generalisability of its results. This type of descriptive data will also provide insight regarding which algorithms and models may not represent certain patient groups adequately.

Fourth, there must be robust processes in the pre-deployment phase to audit model outputs and ensure equal algorithmic performance for diverse patient populations. Sensitivity analyses evaluating algorithmic performance in subgroups can identify biased models with inequitable outcomes. Pre-emptive efforts to adjust models before deployment and mass dissemination protect marginalised subpopulations from inequitable outcomes and can also have cost-saving implications—the excess cost of racial health disparities in the USA is estimated at ~\$230 billion over a 4-year period.[86] For effective impact, the definition of 'inequitable outcome' set by regulators must be fair, clear, specific and quantifiable. Ongoing surveillance in the post-deployment setting is also imperative to monitor for unintended consequences of AI/ML and confirm unbiased algorithmic performance in actuality. Of note, access to training data and prediction methodologies of most large-scale AI/ML algorithms is frequently restricted, thus limiting independent efforts to assess for algorithmic biases and how they may have arisen.[11] This reality underscores the importance of deidentified open-access data sharing in accordance with FAIR[87] data principles—findability, accessibility, interoperability and reusability—which could be highly instrumental in promoting health equity by providing insight into which AI/ML algorithms could perpetuate and/or exacerbate disparities.

Finally, combining AI/ML models with physician clinical decision-making—that is, an augmented intelligence approach with a physician-in-the-loop configuration—may be beneficial in generating ethical and equitable AI/ML tools.[88] Augmented intelligence may be of bidirectional utility as AI/ML models can standardise approaches where considerable provider variability exists while physician interaction can help limit biases that may arise from these tools. However, this approach must be done with careful consideration as biases can also arise from physician interaction with prediction models including automation bias (over-reliance on prediction models), feedback loops, dismissal bias (conscious or unconscious desensitisation) and allocation discrepancy.[89]

While these efforts can minimise bias and create more ethical AI tools, they do not serve as substitutes for repairing medical mistrust[90 91] and certainly do not obviate the structural changes needed to build a more equitable health system.[92–94]

CONCLUSIONS

We describe five themes to illustrate how AI/ML can lead to inequities in gastroenterology/hepatology, examples of the impact on health equity and several potential actionable solutions to ensure equity in AI. By the year 2045, white individuals will comprise less than 50% of the US population, thus this work is critical as AI/ML becomes more common globally and the USA becomes more diverse.[95]

Our primary limitation was the inability to measure or quantify inequities in each clinical example provided. Though each example provided relates directly to a major theme of mechanism of inequities in AI/ML, the degree to which each specific example led to bias cannot be directly measured. In addition, we did not have access to all of the model information used to develop the algorithms discussed nor to robust cost information that could enable a review of cost implications of current AI approaches. This fact highlights the importance of transparency to enable researchers' access to data and inputs included in each algorithm to advance equity. Lastly, the examples provided in this paper are not an exhaustive list but rather focus on strong and relevant illustrations of how prediction models and AI/ML algorithms in gastroenterology and hepatology can lead to biased systems and inequitable health outcomes.

There are several key strengths of this paper. First, we provide clear and actionable solutions to address health equity in AI/ML that can be used by researchers and clinicians alike. Second, we provide concise themes that illustrate how AI/ML can lead to health inequity in gastroenterology matched to specific examples. Our overarching goal is to increase attention to an important potential downside of AI as its use becomes more prevalent and pervasive in the fields of gastroenterology and hepatology.

Here, we adapt a framework to consider equity in AI/ML algorithms used in gastroenterology/hepatology and a platform for discussion around an increasingly relevant topic. In other fields of medicine, we have started to reassess prediction models and algorithms and incorporate a health equity lens. The field of gastroenterology and hepatology has already taken a leading role in clinical applications for AI in medicine, and it is therefore especially important that, as a field, we take a leading role in ensuring that equity considerations are emphasised. This framework will help gastroenterology/hepatology researchers and clinicians prioritise equity in AI/ML development, implementation, and evaluation so that we can give every patient an opportunity to benefit from the technological advances that the future brings.

Twitter Tyler M Berzin @tberzin

Acknowledgements The authors thank Danielle Duffy for graphic design assistance.

Contributors EU-A and AA-Y are joint first authors and contributed equally to this paper. TMB and FPM supervised this project and conceived the original idea. MG critically reviewed the project and provided AI/ML expertise. EU-A, AA-Y, MG, TMB and FPM have approved this version of the manuscript for publication.

Funding FPM is supported by the UCLA Jonsson Comprehensive Cancer Center and the Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program. AA-Y is supported by the National Cancer Institute (award number P50 CA244433) and the Trefler Foundation via MGH Cancer Center.

Competing interests TMB is a consultant for Wision AI, Docbot AI, Medtronic and Magentiq Eye. FPM is a consultant for Medtronic and receives research funding from Exact Sciences. AA-Y receives research funding from Pfizer and Exact Sciences, and consulting fees from Janssen Pharmaceuticals.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Eugenia Uche-Anya http://orcid.org/0000-0003-0261-6573
Tyler M Berzin http://orcid.org/0000-0002-4364-6210

REFERENCES

1 Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2:230–43.
2 Shen L, Margolies LR, Rothstein JH, *et al*. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep* 2019;9:12495.
3 Burdick H, Lam C, Mataraso S, *et al*. Prediction of respiratory decompensation in Covid-19 patients using machine learning: the ready trial. *Comput Biol Med* 2020;124:103949.
4 Nordling L. A fairer way forward for AI in health care. *Nature* 2019;573:S103–5.
5 Pannala R, Krishnan K, Melson J, *et al*. Artificial intelligence in gastrointestinal endoscopy. *VideoGIE* 2020;5:598–613.
6 Le Berre C, Sandborn WJ, Aridhi S, *et al*. Application of artificial intelligence to gastroenterology and hepatology. *Gastroenterology* 2020;158:76–94.
7 Chen B, Garmire L, Calvisi DF, *et al*. Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* 2020;17:238–51.
8 Ahn JC, Connell A, Simonetto DA, *et al*. Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology* 2021;73:2546–63.
9 FDA. Artificial intelligence and machine learning in software as a medical device, 2021. Available: https://wwwfdagov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device#transforming
10 Obermeyer Z, Nissan R, Stern M. *Algorithmic bias Playbook*. University of Chicago Booth, Center for Applied Artificial Intelligence, 2021.
11 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
12 Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019;21:E167–79.
13 Braveman P, Gruskin S. Defining equity in health. *J Epidemiol Community Health* 2003;57:254–8.
14 Chen IY, Pierson E, Rose S, *et al*. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci* 2021;4:123–44.
15 Konkel L. Racial and ethnic disparities in research studies: the challenge of creating more diverse cohorts. *Environ Health Perspect* 2015;123:A297–302.
16 Vollmer S, Mateen BA, Bohner G, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;1:l6927.
17 Polit DF, Beck CT. Generalization in quantitative and qualitative research: myths and strategies. *Int J Nurs Stud* 2010;47:1451–8.
18 Noseworthy PA, Attia ZI, Brewer LC, *et al*. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol* 2020;13:e007988.
19 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
20 Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med* 2020;383:874–82.
21 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46:205–11.
22 Swager A-F, Tearney GJ, Leggett CL, *et al*. Identification of volumetric laser endomicroscopy features predictive for early neoplasia in Barrett's esophagus using high-quality histological correlation. *Gastrointest Endosc* 2017;85:918–26.
23 van der Sommen F, Zinger S, Curvers WL, *et al*. Computer-Aided detection of early neoplastic lesions in Barrett's esophagus. *Endoscopy* 2016;48:617–24.
24 van der Sommen F, Klomp SR, Swager A-F, *et al*. Predictive features for early cancer detection in Barrett's esophagus using volumetric laser endomicroscopy. *Comput Med Imaging Graph* 2018;67:9–20.
25 Ebigbo A, Mendel R, Probst A, *et al*. Real-Time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. *Gut* 2020;69:615–6.
26 Hashimoto R, Requa J, Dao T, *et al*. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest Endosc* 2020;91:1264–71.
27 Islami F, DeSantis CE, Jemal A. Incidence trends of esophageal and gastric cancer subtypes by race, ethnicity, and age in the United States, 1997-2014. *Clin Gastroenterol Hepatol* 2019;17:429–39.
28 Chen S, Zhou K, Yang L, *et al*. Racial differences in esophageal squamous cell carcinoma: incidence and molecular features. *Biomed Res Int* 2017;2017:1204082.
29 Zhang Y. Epidemiology of esophageal cancer. *World J Gastroenterol* 2013;19:5598–606.
30 Ohmori M, Ishihara R, Aoyama K, *et al*. Endoscopic detection and differentiation of esophageal lesions using a deep neural network. *Gastrointest Endosc* 2020;91:301–9.
31 Horie Y, Yoshio T, Aoyama K, *et al*. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc* 2019;89:25–32.
32 Cai S-L, Li B, Tan W-M, *et al*. Using a deep learning system in endoscopy for screening of early esophageal squamous cell carcinoma (with video). *Gastrointest Endosc* 2019;90:745–53.
33 Tokai Y, Yoshio T, Aoyama K, *et al*. Application of artificial intelligence using convolutional neural networks in determining the invasion depth of esophageal squamous cell carcinoma. *Esophagus* 2020;17:250–6.
34 Guo L, Xiao X, Wu C, *et al*. Real-Time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). *Gastrointest Endosc* 2020;91:41–51.
35 Kohli A, Holzwanger EA, Levy AN. Emerging use of artificial intelligence in inflammatory bowel disease. *World J Gastroenterol* 2020;26:6923–8.
36 Waljee AK, Lipson R, Wiitala WL, *et al*. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis* 2017;24:45–53.
37 Waljee AK, Liu B, Sauder K, *et al*. Predicting Corticosteroid-Free biologic remission with Vedolizumab in Crohn's disease. *Inflamm Bowel Dis* 2018;24:1185–92.
38 Gan RW, Sun D, Tatro AR, *et al*. Replicating prediction algorithms for hospitalization and corticosteroid use in patients with inflammatory bowel disease. *PLoS One* 2021;16:e0257520.
39 Aniwan S, Harmsen WS, Tremaine WJ, *et al*. Incidence of inflammatory bowel disease by race and ethnicity in a population-based inception cohort from 1970 through 2010. *Therap Adv Gastroenterol* 2019;12:1756284819827692.
40 Afzali A, Cross RK. Racial and ethnic minorities with inflammatory bowel disease in the United States: a systematic review of disease characteristics and differences. *Inflamm Bowel Dis* 2016;22:2023–40.
41 Ng SC, Shi HY, Hamidi N, *et al*. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet* 2017;390:2769–78.

42 Dos Santos Marques IC, Theiss LM, Wood LN, *et al*. Racial disparities exist in surgical outcomes for patients with inflammatory bowel disease. *Am J Surg* 2021;221:668–74.
43 Barnes EL, Loftus EV, Kappelman MD. Effects of race and ethnicity on diagnosis and management of inflammatory bowel diseases. *Gastroenterology* 2021;160:677–89.
44 Spann A, Yasodhara A, Kang J, *et al*. Applying machine learning in liver disease and transplantation: a comprehensive review. *Hepatology* 2020;71:1093–105.
45 Merion RM, Sharma P, Mathur AK, *et al*. Evidence-Based development of liver allocation: a review. *Transpl Int* 2011;24:965–72.
46 Ahearn A. Ethical dilemmas in liver transplant organ allocation: is it time for a new mathematical model? *AMA J Ethics* 2016;18:126–32.
47 Cholongitas E, Marelli L, Kerry A, *et al*. Female liver transplant recipients with the same GFR as male recipients have lower MELD scores--a systematic bias. *Am J Transplant* 2007;7:685–92.
48 Mindikoglu AL, Regev A, Seliger SL, *et al*. Gender disparity in liver transplant waiting-list mortality: the importance of kidney function. *Liver Transpl* 2010;16:1147–57.
49 Allen AM, Heimbach JK, Larson JJ, *et al*. Reduced access to liver transplantation in women: role of height, MELD exception scores, and renal function underestimation. *Transplantation* 2018;102:1710–6.
50 Martin P, DiMartini A, Feng S, *et al*. Evaluation for liver transplantation in adults: 2013 practice guideline by the American association for the study of liver diseases and the American Society of transplantation. *Hepatology* 2014;59:1144–65.
51 Massie AB, Caffo B, Gentry SE, *et al*. MELD exceptions and rates of waiting list outcomes. *Am J Transplant* 2011;11:2362–71.
52 Nephew LD, Goldberg DS, Lewis JD, *et al*. Exception points and body size contribute to gender disparity in liver transplantation. *Clin Gastroenterol Hepatol* 2017;15:1286–93.
53 Petrick JL, Braunlin M, Laversanne M, *et al*. International trends in liver cancer incidence, overall and by histologic subtype, 1978-2007. *Int J Cancer* 2016;139:1534–45.
54 Singal AG, El-Serag HB. Hepatocellular carcinoma from epidemiology to prevention: translating knowledge into practice. *Clin Gastroenterol Hepatol* 2015;13:2140–51.
55 Locke JE, Shelton BA, Olthoff KM, *et al*. Quantifying Sex-Based disparities in liver allocation. *JAMA Surg* 2020;155:e201129.
56 Nitski O, Azhie A, Qazi-Arisar FA, *et al*. Long-Term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. *Lancet Digit Health* 2021;3:e295–305.
57 Wahid NA, Rosenblatt R, Brown RS. A review of the current state of liver transplantation disparities. *Liver Transpl* 2021;27:434–43.
58 Ahmed S, Nutt CT, Eneanya ND, *et al*. Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. *J Gen Intern Med* 2021;36:464–71.
59 Corley DA, Jensen CD, Marks AR, *et al*. Adenoma detection rate and risk of colorectal cancer and death. *N Engl J Med* 2014;370:1298–306.
60 Corley DA, Jensen CD, Marks AR. Can we improve adenoma detection rates? A systematic review of intervention studies. *Gastrointest Endosc* 2011;74:656–65.
61 Repici A, Badalamenti M, Maselli R, *et al*. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* 2020;159:512–20.
62 Hassan C, Spadaccini M, Iannone A, *et al*. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointest Endosc* 2021;93:77–85.
63 Rex DK, Boland CR, Dominitz JA, *et al*. Colorectal cancer screening: recommendations for physicians and patients from the U.S. Multi-Society Task force on colorectal cancer. *Gastrointest Endosc* 2017;86:18–33.
64 Zhou G, Xiao X, Tu M, *et al*. Computer aided detection for laterally spreading tumors and sessile serrated adenomas during colonoscopy. *PLoS One* 2020;15:e0231880.
65 Jackson CS, Vega KJ. Higher prevalence of proximal colon polyps and villous histology in African-Americans undergoing colonoscopy at a single equal access center. *J Gastrointest Oncol* 2015;6:638–43.
66 Nouraie M, Hosseinkhah F, Brim H, *et al*. Clinicopathological features of colon polyps from African-Americans. *Dig Dis Sci* 2010;55:1442–9.
67 Thornton JG, Morris AM, Thornton JD, *et al*. Racial variation in colorectal polyp and tumor location. *J Natl Med Assoc* 2007;99:723–8.
68 Devall M, Sun X, Yuan F, *et al*. Racial disparities in epigenetic aging of the right vs left colon. *J Natl Cancer Inst* 2020. doi:10.1093/jnci/djaa206. [Epub ahead of print: 30 Dec 2020].
69 Nouraie M, Ashktorab H, Atefi N, *et al*. Can the rate and location of sessile serrated polyps be part of colorectal cancer disparity in African Americans? *BMC Gastroenterol* 2019;19:77.
70 Ashktorab H, Delker D, Kanth P, *et al*. Molecular characterization of sessile serrated Adenoma/Polyps from a large African American cohort. *Gastroenterology* 2019;157:572–4.
71 DeSantis CE, Miller KD, Goding Sauer A, *et al*. Cancer statistics for African Americans, 2019. *CA Cancer J Clin* 2019;69:211–33.
72 Fedewa SA, Flanders WD, Ward KC, *et al*. Racial and ethnic disparities in interval colorectal cancer incidence: a population-based cohort study. *Ann Intern Med* 2017;166:857–66.
73 Almario CV, May FP, Ponce NA, *et al*. Racial and ethnic disparities in colonoscopic examination of individuals with a family history of colorectal cancer. *Clin Gastroenterol Hepatol* 2015;13:1487–95.
74 Benarroch-Gampel J, Sheffield KM, Lin Y-L, *et al*. Colonoscopist and primary care physician supply and disparities in colorectal cancer screening. *Health Serv Res* 2012;47:1137–57.
75 Laiyemo AO, Doubeni C, Pinsky PF, *et al*. Race and colorectal cancer disparities: health-care utilization vs different cancer susceptibilities. *J Natl Cancer Inst* 2010;102:538–46.
76 Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, *et al*. Contribution of screening and survival differences to racial disparities in colorectal cancer rates. *Cancer Epidemiol Biomarkers Prev* 2012;21:728–36.
77 Goldberg DS, Taddei TH, Serper M, *et al*. Identifying barriers to hepatocellular carcinoma surveillance in a national sample of patients with cirrhosis. *Hepatology* 2017;65:864–74.
78 Singal AG, Li X, Tiro J, *et al*. Racial, social, and clinical determinants of hepatocellular carcinoma surveillance. *Am J Med* 2015;128:90.e1–7.
79 Riall TS, Townsend CM, Kuo Y-F, *et al*. Dissecting racial disparities in the treatment of patients with locoregional pancreatic cancer: a 2-step process. *Cancer* 2010;116:930–9.
80 Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun* 2020;11:5131.
81 Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol* 2019;25:1666–83.
82 Chen H, Sung JJY. Potentials of AI in medical image analysis in gastroenterology and hepatology. *J Gastroenterol Hepatol* 2021;36:31–8.
83 Suresh H, Guttag J. Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle. In: *Mit case studies in social and ethical responsibilities of computing*, 2021.
84 Gebru T, Morgenstern J, Vecchione B. *Datasheets for datasets. communications of the ACM*, 2021: 86–92.
85 Mitchell M, Wu S, Zaldivar A. *Model cards for model reporting. Proceedings of the conference on Fairness, accountability, and transparency*, 2019: 220–9.
86 LaVeist TA, Gaskin D, Richard P. Estimating the economic burden of racial health inequalities in the United States. *Int J Health Serv* 2011;41:231–8.
87 Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al*. The fair guiding principles for scientific data management and stewardship. *Sci Data* 2016;3:160018.
88 Crigger E, Reinbold K, Hanson C, *et al*. Trustworthy augmented intelligence in health care. *J Med Syst* 2022;46:12.
89 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
90 DeCamp M, Tilburt JC. Why we cannot trust artificial intelligence in medicine. *Lancet Digit Health* 2019;1:e390.
91 Ryan M. In AI we trust: ethics, artificial intelligence, and reliability. *Sci Eng Ethics* 2020;26:2749–67.
92 Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med* 2018;378:981–3.
93 Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull World Health Organ* 2020;98:245–50.
94 Baron RJ, Khullar D. Building trust to promote a more equitable health care system. *Ann Intern Med* 2021;174:548–9.
95 Bahrampour T, Mellnik T. *Census data shows widening diversity; number of white people falls for first time*. Washington Post, 2021.