Multi-disciplinary fairness considerations in machine learning for clinical trials

Isabel Chien
ic390@cam.ac.uk
University of Cambridge
Cambridge, UK

Nina Deliu
nina.deliu@mrc-bsu.cam.ac.uk
MRC Biostatistics Unit, University of
Cambridge
Cambridge, UK
MEMOTEF, Sapienza University of
Rome
Rome, IT

Richard E. Turner
ret26@cam.ac.uk
University of Cambridge
Cambridge, UK

Adrian Weller
aw665@cam.ac.uk
University of Cambridge
Cambridge, UK
The Alan Turing Institute
London, UK

Sofia S. Villar
sofia.villar@mrc-bsu.cam.ac.uk
MRC Biostatistics Unit, University of
Cambridge
Cambridge, UK

Niki Kilbertus
niki.kilbertus@tum.de
Technical University of Munich
Munich, Germany
Helmholtz Munich
Munich, Germany

ABSTRACT

While interest in the application of machine learning to improve healthcare has grown tremendously in recent years, a number of barriers prevent deployment in medical practice. A notable concern is the potential to exacerbate entrenched biases and existing health disparities in society. The area of fairness in machine learning seeks to address these issues of equity; however, appropriate approaches are context-dependent, necessitating domain-specific consideration. We focus on clinical trials, i.e., research studies conducted on humans to evaluate medical treatments. Clinical trials are a relatively under-explored application in machine learning for healthcare, in part due to complex ethical, legal, and regulatory requirements and high costs. Our aim is to provide a multi-disciplinary assessment of how fairness for machine learning fits into the context of clinical trials research and practice. We start by reviewing the current ethical considerations and guidelines for clinical trials and examine their relationship with common definitions of fairness in machine learning. We examine potential sources of unfairness in clinical trials, providing concrete examples, and discuss the role machine learning might play in either mitigating potential biases or exacerbating them when applied without care. Particular focus is given to adaptive clinical trials, which may employ machine learning. Finally, we highlight concepts that require further investigation and development, and emphasize new approaches to fairness that may be relevant to the design of clinical trials.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Health informatics**.

KEYWORDS

clinical trials, adaptive clinical trials, health informatics, machine learning for healthcare

1 INTRODUCTION

Recent advancements in the field of machine learning (ML) have fueled considerable excitement regarding its potential to change the landscape of healthcare and biomedical research globally [184, 200]. While much of current research has focused on the application of ML to benefit clinical practice, ML may also offer improvements to clinical research [267, 275]. Clinical research refers to the study of the safety and efficacy of medical, surgical, or behavioral interventions that are intended for human use [30]. A common tool of clinical research is the clinical trial (CT): experiments conducted to evaluate the impact of treatments or interventions on human subjects, ranging from early-phase dose-finding trials to confirmatory studies. Randomized controlled trials (RCTs) are considered the "gold standard" for evaluating the effectiveness of experimental interventions [6, 234], due to their strong statistical guarantees [30, 209]. However, the traditional fixed design allows no flexibility for beneficial alterations. As a result, interest is growing in adaptive designs (ADs), wherein results from an ongoing clinical trial may be incorporated into trial procedure changes [51, 187].

Researchers and commercial entities have proposed ML-driven improvements to the design of clinical trials, but extensive research

and adoption in practice have been limited [267]. A likely explanation is that clinical trials are subject to a complex mixture of ethical, legal, and regulatory considerations, and, as yet, there exists limited guidance for the inclusion of ML [92, 267]. The design of clinical trials remains the subject of ongoing ethical discussions, with some criticizing the morality of RCTs [179, 212], and others detailing possible issues surrounding adaptive designs [30, 189, 248]. To complicate matters further, ethical principles currently considered when assessing clinical research studies may be insufficient to address additional concerns stemming from the inclusion of ML.

Nevertheless, ML has been proposed to assist with clinical trials planning, participant management, data collection, and analysis [267]. Deep learning models have been developed for patient-trial matching, using electronic health records (EHR) data and trial eligibility criteria to recommend suitable patients for certain trials [105, 287]. Such systems have performed well in practice; a study found that the IBM Watson for Clinical Trial matching system, used to match breast cancer patients with systemic therapy trials, increased average monthly enrollment by 80% over the 18 months following implementation [100]. From a more theoretical perspective, reinforcement learning (RL) and multi-armed bandit (MAB) methods have been considered for modelling safe, effective doses in adaptive dose-finding trials [15, 148, 224].

While ML has potential to improve the welfare and fairness across participants in clinical trials, there is a risk that ML exacerbates existing inequalities or introduces new disparities and biases [81, 191, 270]. ML model outputs are dependent on their training datasets, which may reflect historical biases [201], thus perpetuating existent disparities in society. Women and racial/ethnic minorities are systematically underrepresented in clinical research, contributing to disproportionately higher incidence of and mortality from cancer [66] and cardiovascular diseases [48], and public health misconceptions [32]. Healthcare algorithms used in practice as well as those considered state-of-the-art for certain clinical prediction tasks have been discovered to exhibit racial [185], gender, and socioeconomic biases [44].

Fairness in machine learning addresses disparities between subpopulations in data-driven decision-making systems to mitigate inequalities. However, the intersection of machine learning, fairness, and healthcare is still relatively under-explored, with many unresolved ethical questions [40]. Existing fairness criteria have been found to be mutually incompatible [191] and conceptions of fairness addressing exploration-exploitation trade-offs require more attention [50]. Approaches to and definitions of fairness are context-dependent, requiring domain-specific consideration. An interdisciplinary approach including ethicists, medical practitioners, social and other quantitative scientists is required to assess appropriate formal fairness measures, interventions, and necessary trade-offs [201, 270]. Bias can arise at every step of the process, including study design, data collection, data analysis, model building, model evaluation, and even deployment [40, 43, 201]. Researchers increasingly emphasize the need for formal rules and regulations that explicitly address issues of fairness in ML for healthcare [92].

In this paper, we aim to contextualize fairness considerations within ML for clinical trials, to provide structure and clarity for those hoping to leverage ML to improve the practice of clinical trials. We consider ML from several angles: its potential to improve

patient welfare and equity, the risks of propagating unfairness and adverse effects, and its potential to facilitate the use of adaptive designs, which could be considered fairer than RCTs. We highlight the following contributions:

(1) After a primer on the design and implementation of clinical trials, we discuss the surrounding ethical considerations and guidelines from an ML perspective and establish relationships with notions of fairness in ML.
(2) We examine potential sources of unfairness in clinical trials alongside concrete examples to illustrate where adaptive designs or ML can be used to facilitate improvements.
(3) We curate a list of opportunities for innovation in ML to address sources of unfairness, providing an overview of current literature and highlighting critical directions for future work.

2 BACKGROUND: CLINICAL TRIALS, ETHICS, AND FAIRNESS

In this section, we provide background for the multi-disciplinary discussion of ethical usage of ML in clinical trials. We characterize the modern practice of clinical trials, from design to implementation. Following that, we review influential documents and regulations that govern clinical research alongside renewed contemporary ethical discussions. Finally, we briefly describe ongoing approaches to and open questions within fairness in ML.

2.1 Structure of Clinical Trials

2.1.1 Clinical Trial Designs. CTs are research studies in which participants are allocated to different treatments in order to evaluate the effectiveness and safety of treatments of interest. They are commonly classified into four phases: dose-finding phase-I trials, testing for safety in a few human subjects; phase-II efficacy evaluation trials; confirmatory phase-III trials, testing for effectiveness; and phase-IV surveillance studies, for long-term effects monitoring [56].

Patient allocation may be randomized, where each participant has a known positive probability of being assigned to a treatment arm, or non-randomized [127] with rule-based approaches [269]. Well-designed RCTs are recognized as the "gold standard" for conducting confirmatory evidence-based evaluation of experimental interventions [6, 234]. In a RCT, participants are randomly allocated to either control arms (standard of care or placebo), or experimental arms [130]. Randomization removes the influence of potential confounders on study outcomes [215, 233], mitigating bias and enabling valid cause-effect estimation [209]. Traditional RCTs are static; key elements (e.g., sample size, randomization probabilities) remain fixed during the course of the trial, with the exception that a trial may be stopped due to significant safety concerns. This may be limiting, as interim analysis may provide enough evidence for stopping the trial earlier for success, lack of efficacy, or skewing allocation towards the most beneficial treatment.

In adaptive designs, key trial characteristics (e.g., treatment doses or allocation probabilities) may be altered during the course of the study based on accumulating data (e.g., treatment responses), according to predefined rules [51, 187]. Examples include: (i) *response-adaptive randomization* (RAR), where allocation ratios may be shifted towards more promising or informative treatments [187]; (ii) *drop the loser*, where inferior treatment arms may be removed [36, 187]; or (iii) *adaptive dose-finding*, where safe (maximum tolerated)

dosages are determined according to rule- or model-based strategies [36, 269]. Multiple adaptations may also be incorporated into a single trial. While adaptive designs can be more efficient, informative, and ethical than traditional RCTs, as they are able to adjust to new information and make better use of limited resources [51, 187], they may create new ethical issues, increase costs, and complicate statistical analysis. For a detailed overview on adaptive designs, we refer the interested reader to [20, 25, 187].

While public interest in adaptive designs seems to rise during healthcare crises, uptake in practice has been minimal, particularly in relation to the methodological/theoretical literature [206, 237]. Common concerns about adaptive designs stem from lack of clarity on the methodology involved in running such trials and applicability of traditional statistical inference methods to interpret trial outcomes [187]. In recent years, proponents of adaptive designs have sought to guide understanding of how they can be successfully implemented and what they can achieve [9, 36, 51, 187, 206], and to pursue developments in machine learning that may support adaptive methods [15, 19, 150, 256, 260, 272, 273].

2.1.2 Clinical Trials in Practice. Proposed CTs are typically outlined in a *clinical trial protocol*, which must then be submitted to regulatory bodies and ethics committees for review and approval before trial implementation [39]. It is a document written by study investigators that delineates the motivations of a trial and the planned methodologies by which a trial will be conducted [39, 205]. To help standardize trial protocols, an international committee published the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) guidelines (2013) [39], as well as a SPIRIT-AI Extension (2020), which addresses interventions that involve AI [205]. The SPIRIT guidelines recommend that a trial protocol include comprehensive details such as (but not limited to) justification for undertaking the trial, eligibility criteria, interventions being studied, safety considerations, trial design, and the statistical analysis plan.

The regulatory bodies and ethics committees responsible for trial approval vary by country. In the United States (US), the Food and Drug Administration (FDA) is responsible for regulatory approval [56] and Institutional Review Boards independently determine ethical approval [83]. In the United Kingdom, trial protocols are submitted to multiple agencies for various ethics and regulatory approvals [1]. Trials are assessed based on the aim of the trial, whether likely benefits outweigh risks, the planned design and analysis of the trial, among other aspects [3, 56]. Treatments are typically approved by regulatory bodies following successful confirmatory phase-III trials [56].

2.2 Ethics of Clinical Trials

2.2.1 Current Ethical Guidelines. Ethical guidelines regarding the conduct of clinical research have been developed for guidance against harming or exploiting patient volunteers, while preserving the integrity of the scientific research. Emanuel et al. [76] have developed a widely-cited framework consisting of seven requirements for assessing the ethics of clinical research studies. We take guidance from these seven requirements [2, 76], with further discussion in Appendix A: (1) value for society, (2) scientific validity, (3) fair subject selection, (4) favorable risk-benefit ratio, (5) independent review, (6) informed consent, and (7) respect for enrolled subjects.

However, these guidelines do not sufficiently address issues arising from the growing uptake of novel trial designs or the use of ML in clinical trials settings, which will inevitably bring about risks and burdens that should be explicitly acknowledged in any trial planning process.

2.2.2 Guidelines for AI/ML in Clinical Trials. Currently, there exists limited guidance on the ethical inclusion of ML in clinical research. Fairness is a key consideration in ethical ML, alongside explainability, privacy, accountability, and transparency. Neither the US nor the European Union (EU) provide any guidance specific to ML in clinical research, although the FDA provides guidelines for regulatory assessment of ML for clinical care [56], and the EU provides guidance on the general development of AI [242]. With respect to clinical trials, international committees have recently published the SPIRIT-AI guidelines for trial protocols [205] and the CONSORT-AI guidelines for trial reporting [161]. While these address the practical inclusion of AI/ML techniques in clinical research, they mention fairness only in passing. Accordingly, researchers have emphasized the need for explicit guidance and recommendations of minimum fairness standards from regulatory bodies [92]. We argue for greater consideration of issues of fairness in the development of ML for healthcare from ideation to implementation.

2.3 Fairness in Machine Learning

The study of fairness in machine learning aims to ensure that ML models do not "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [88]. Individuals or groups are typically characterized via certain (demographic) attributes associated with sub-populations who are vulnerable to harm due to structural biases [171], such as age or race. These characteristics are legally termed *sensitive* or *protected* attributes [17]. Here, we briefly review existing notions of fairness in situations where machine learning models inform or make decisions that affect the livelihood or well-being of humans and summarize challenges with these notions. We defer an introductory overview of existing fairness definitions and approaches to achieve them in practice to Appendix B.

There are several popular perspectives seeking to understand how discrimination of data-driven decision-making can be formalized. Just like ethicists, political philosophers, legal scholars, and social scientists still grapple with what is fair, just, or equitable, there is an ongoing debate about the most reasonable formal definitions of fairness in ML and which ethical principles to consider [18, 26, 87, 95, 112]. In practice, often only statistical *group fairness* notions can be operationalized well enough in a task-agnostic manner to suit real-world application. However, beyond incompatibility results between these definitions [49, 142]–and consequently the difficulty of choosing one out of many group fairness definitions–a comprehensive empirical examination of the impact of penalizing group fairness violations on a variety of measures of model performance and group fairness metrics on various clinical datasets and prediction tasks found that such procedures lead to near-universal performance degradation along multiple dimensions. This indicates that group fairness measures are often not an effective solution for handling unfairness in healthcare settings [194]. On the other hand, *individual fairness* suffers from the required normative choice of a "similarity metric" on individuals, which is difficult to

formalize even in well-studied contexts. While causal definitions are conceptually closest to legal notions of fairness [164], there still remains fundamental disagreement about the most applicable concepts of fairness in particular settings [117, 191]. Causal definitions require strong assumptions on the underlying data generating mechanism [137, 144, 164] and are thus difficult to operationalize. Since there is no universal technical solution to fairness in ML, the path forward is to involve domain experts, affected stakeholders, and to recognize social context, including the potential for institutionalized discrimination, and examine each situation holistically [24, 194, 220].

3 POSSIBLE SOURCES OF UNFAIRNESS IN CLINICAL TRIALS

While ethical guidelines attempt to advise in a principled and humane approach towards clinical research, there still continues considerable debate regarding the construction of clinical trials. With respect to the seven principles of ethical clinical studies (Section 2.2.1), we delineate possible sources of unfairness to expose where further advancement of trial designs and implementation of fair ML can contribute towards more ethical clinical research.

3.1 Subject Participation

Women, racial/ethnic minorities, and other under-served groups have been historically under-represented in CTs [168]. Disparities have been identified in the study of cardiovascular disease [48, 109, 202], oncology [46, 73], and acquired immunodeficiency syndrome (AIDS) [98, 225]; these particularly affect women of racial/ethnic minorities [98, 159]. While most CTs have been implemented in relatively wealthy regions such as North America and Western Europe [239], the pharmaceutical industry has driven globalization of clinical research, particularly to developing countries [60, 223]. Although this may improve diversity, it also requires renewed consideration of ethics and fairness [60, 223, 239].

The value of inclusion has been widely demonstrated, with significant survival benefits [278]. Disparities in medical knowledge have long been criticized, with many questioning how the "white male [came] to be the prototype of the human research subject" [38, 59, 72]. This may result in under-studied groups receiving ineffective or harmful treatments (or none at all) limited access to hospital care [227]. For example, women experience a notably higher incidence of adverse drug reaction and toxicity than men, but few trials explore sex differences [246, 289, 290]. Under-representation can also affect statistical trial outcomes, diminishing knowledge generalizability [120, 133, 210, 266]. Unfair under-representation may occur during the process of subject selection, recruitment, and enrollment in a clinical trial. It may arise from explicit exclusion of certain subgroups (Section 3.1.1), or under-representation caused by socio-economic disparities, discriminatory practices, or communication barriers (Section 3.1.2).

3.1.1 Categorical Subject Exclusion. International ethical guidelines for conducting human medical research have asserted that subjects of research should be selected such that burdens and benefits are equitably distributed and exclusion of groups should be justified [11, 76, 84, 154]. However, studies have found that inclusion/exclusion criteria are often restrictive and unjustified in their exclusion of under-served groups [29, 131, 202, 229]. Although some exclusion may be sensible (e.g., people without prostates from

prostate cancer studies), researchers have identified explicitly codified exclusion of certain racial/ethnic groups [202], women, the elderly, and those with co-morbidities without good reason [109, 131, 229]. Eligibility criteria may also indirectly exclude sub-populations; some studies excluded participants on the basis of language and birthplace [202] and others on the basis of co-morbidities that are more prevalent amongst certain racial/ethnic groups [114]. The inclusion of vulnerable individuals (e.g., pregnant women [21], children, or people from low-income countries) has required particular justification [33], often leading to their exclusion.

> **Case 1: Participation of pregnant women in clinical research.** Pregnant and lactating women remain under-represented in clinical research, despite advocacy from international organizations [84]. Clinical testing of experimental vaccines has historically excluded pregnant women, due to potential harm to the fetus, physiological complexity of pregnancy, and legal or regulatory complications [219]. Consequently, they and their infants are often unfairly deprived of protections against potentially catastrophic infections. For example, although generally asymptomatic, the Zika virus can produce congenital malformation syndrome and pediatric neuro-developmental abnormalities [218]; an effective Zika vaccine would be most beneficial to unborn infants. Vaccine-preventable infections are also a leading cause of morbidity in pregnant women [78, 79]. Pregnant women have also been excluded from clinical trials during a recent Ebola outbreak in 2018 [34, 219] and the current COVID-19 pandemic [169, 213, 238], despite mounting evidence that pregnant woman may have a higher risk of COVID-19 complications [255].

3.1.2 Under-representation. Even in the absence of explicit categorical exclusion, under-representation may result from i) unequal access to care due to socio-economic and geographical differences, ii) implicit physician biases, and iii) communication issues between patients and investigators, among other causes [53, 82, 180, 183]. Subgroups with historically lower financial resources are consistently under-represented in cancer clinical trials, likely owing to high costs of frequent travel to trial sites, inability to take time away from work, or caution against the financial impacts of adverse effects, compounding with already high routine care costs [53, 183]. People of color have also historically had inadequate access to medical care; minorities are more likely to receive care at under-resourced hospitals [104] and are less likely to have health insurance, which is often a pre-requisite for medical facilities and clinical trials [82].

Physicians are often responsible for determining who to approach for enrollment. Those who harbor biases may believe that minorities are less likely to remain adherent to treatment recommendations [214, 251], or likely to reject trial offers [102]. However, research has shown that minorities are comparably willing to participate in clinical trials, but are approached far less [268]. Many clinical trials recruit on a first-come, first served basis; those with wealth, power, and connections are thus more likely to receive scarce resources, a particular challenge in unforeseen catastrophes such as the COVID-19 pandemic [123]. Communication issues affecting economically and educationally disadvantaged individuals, such as language-related factors, lack of access to technology, or low levels of health literacy [65, 235], may also undermine fair subject selection, particularly in the absence of facilitating instruments (e.g., interpreters).

Case 2: Women in AIDS studies. Perhaps due to the historical stereotype that AIDS primarily afflicts gay men, its impact on women has been under-studied, with women excluded from studies or included in numbers too small to yield meaningful outcomes. Although infections are increasing most rapidly amongst women, the number participating in clinical trials of AIDS drugs, which provide an important source of first-rate medical care and access to experimental treatments, lags behind expectations [225]. As a result, lower case survival rates have been observed in women diagnosed with AIDS. Studies involving women have focused on the reduction or prevention of transmission of human immunodeficiency virus (HIV) from a pregnant woman to a fetus or newborn, not on female-specific manifestations of the disease. Recent efforts have begun to address these concerns, arguing that what is known about the natural history of the diseases and their treatment in men may be inapplicable to women [225].

3.2 Treatment Allocation

A critical component of a trial design is the mechanism by which participants are allocated to treatment arms. In the absence of scientific and ethical justification, randomized schemes are preferable in most circumstances, as they protect against allocation bias and facilitate equal baseline distribution of prognostic factors between the compared groups [188]. As reported in the CIOMS International Ethical Guidelines, RCTs in particular equalize the foreseeable benefits and risks of participation [84] and also provide recognized scientific and statistical superiority [209, 234]. However, even in RCTs, issues of unfairness may still arise, either due to their inherent characteristics or to violations of ethical principles.

3.2.1 Randomized Controlled Trials. The moral justification of clinical trials relies on the principle of *clinical equipoise*, wherein researchers must possess "genuine uncertainty" regarding the benefits of new therapies over existing standard practices [85, 158]. However, RCTs typically investigate treatments that show promise in pilot trials, but are usually conducted with a fixed equal allocation ratio, uniformly distributing benefits and burden. While successful small-scale trials are not considered sufficient evidence of treatment effectiveness by regulatory agencies, as they often lack statistical power, they may influence an investigator's uncertainty. Preliminary evidence from pilot trials (and from uncontrolled trials [69]) may provide investigators with reason to believe in the success of an intervention, especially if the treatment effect is large, violating clinical equipoise. Adaptive designs with unequal [172] or adaptive allocation ratios [14, 193] (based on treatment success or patient characteristics) may be more ethical alternatives, given the responsibility of practitioners to minimize the number of people given unsafe or inferior treatments (or placebos) [25, 187].

Case 3: Ethics of RCTs. The ACTG076 trial was conducted to examine whether zidovudine (AZT) could reduce vertical maternal-to-infant HIV transmission, compared against a placebo [57]. A conventional randomization design was used to ensure equal allocation of pregnant women, with 239 receiving AZT and 238 receiving placebo. However, 16 newborns in the AZT group and 52 in the placebo group had at least one HIV-positive culture, with a higher (>3x) HIV transmission in infants on placebo treatment. Subsequent researchers found that a suitable AD could have saved 11 newborns without significant loss of efficiency [27, 279].

Clinical equipoise is used to justify the contradiction between physicians' therapeutic obligation to treat patients with the most beneficial possible methods and the true goal of clinical research: collecting evidence [173]. This mindset has been criticized as contributing to *therapeutic misconception*, where individuals believe that the primary goal of a clinical trial is to improve their outcomes, while the true goal of a trial is to produce experimentally-validated knowledge, where scientific validity is prioritized over patient outcomes [110, 172]. Adaptive designs may provide a way to improve patient outcomes within clinical trials, while maintaining scientific validity, thereby reducing the impacts of therapeutic misconception [172].

3.2.2 Adaptive designs: benefits and ethical controversies. In some cases, the flexible and efficient nature of adaptive designs may offer a corrective to certain ethical difficulties of traditional RCTs [36, 258], resulting in more patient benefits [25, 152, 187]. However, adaptive designs generate their own ethical concerns [165] and may propagate inequalities. For example, clinical equipoise may be violated when data from the first group of study participants is evaluated [145, 216]. In RAR designs where allocation probabilities are skewed in favor of more promising treatment(s), it may be difficult to fully blind investigators to treatment arms. As enrollment progresses, investigators may recognize that more patients are being assigned to a certain arm and infer the superiority of the associated treatment [145, 216]. Sequential adaptation may also lead to different treatment of patients depending on time of enrollment [153]. The later a patient joins the study in relation to the target sample size, the higher their chance of receiving a superior treatment under the changing randomization scheme, leading to an unfair distribution of risk and benefit across participants.

Case 4: Benefits of Adaptive Designs. Lee et al. [151] argue that equipoise may be violated during COVID-19 RCTs that study treatments that are expected to work. Following an interim analysis in the adaptive ACTT-1 trial for evaluating the experimental remdesivir treatment for COVID-19 [23], which found that patients treated with remdesivir had a 31% faster time to recovery than those who received placebo, the National Institute of Allergy and Infectious Diseases (NIAID) chose to offer patients in the placebo group the opportunity to receive remdesivir. Although this decision was sharply criticized, it was ethically justifiable, as at the time the NIAID chose to stop the trial, it was clear the treatment was more effective than placebo [176]. Adaptive designs can expose fewer patients to ineffective and/or toxic interventions, ensuring a fairer distribution of superior treatments [189].

Trial adaptations based on interim results are not guaranteed to be beneficial, as both partial and final results may be inaccurate and biased in both magnitude and direction [207, 256]. The sequential dependence induced by adaptive designs can lead to problems in confirmatory trials with frequentist errors controls [64, 256], exposing participants to inferior treatments and invalidating study results. Additionally, while adaptations should be pre-defined in a trial protocol, Park et al. [189] caution that investigators may make changes post-hoc, based on results, introducing investigator-driven bias, although such issues could be mitigated with an independent monitoring committee. Legocki et al. [153] discuss concerns of informed consent due to the difficulties of clearly explaining the structure of adaptive designs to potential participants [153]. Adoption of adaptive designs is also hindered by perceived technical and logistical

complexity [51]; current works have sought to clarify methodology to improve accessibility to a wider audience [51, 187, 206].

3.3 Throughout a Trial

Sources of unfairness may arise at trial planning or subject selection, but extend throughout trial implementation and generate post-trial consequences.

3.3.1 Differing Patient Risk Preferences. CT participants may operate with distinct risk preferences, defined in the economics and psychology literature as the "extent to which people are willing to take on risk" [41]. In clinical research, these are described as *benefit-risk preferences*, where patients may find varying severity of side effects or consequences to be acceptable given potential benefits [106, 124, 163, 204]. For example, a survey of older Americans found that they were willing to accept considerable risk–a 1-year risk of over 30% of death or permanent severe disability from stroke–in exchange for an Alzheimer's disease treatment that would prevent disease progression beyond a mild stage [107].

Case 5: A study examining counseling for post-mastectomy breast cancer patients. An RCT sought to evaluate the effectiveness of counselling versus no counselling for participants undergoing mastectomy for breast cancer [35]. However, willingness to engage in counselling would have had a significant impact on the overall success of counselling, confounding treatment effects. This would lead to invalid results; those who preferred to enter counselling may have felt improvements as a result, but those effects would be obfuscated by those who did want to engage in counselling and inevitably did not experience improvements. In response, Brewin and Bradley [35] propose a preference-based trial design, where patients are allocated to their treatment preferences (unless they have none, in which case they are randomized) to mitigate these confounders.

Researchers, regulators, and private companies increasingly agree upon the importance of incorporating patient preferences into clinical research [228, 231, 250]. Patient risk preferences provide insight into the types of treatments patients prefer, the consequences they may tolerate, and what they consider to be meaningful clinical benefit [174, 231]. Elicitation of risk preferences may aid in patient recruitment, as uncertainties regarding risk prevent some from participating in CTs [183]. For example, a study of perceptions of cardiovascular CTs found that women perceived greater harm from trial participation than men, contributing to their underrepresentation in cardiovascular CTs [70].

Many patients also express a dislike of randomization, distrust of researchers, or strong treatment preferences [243, 245], causing them to refuse participation, negatively impacting the generalizability of a study to the outside population. Patients who agree to randomization but receive a non-preferred treatment may suffer resentful demoralization [35], resulting in poor treatment compliance or dropout, or a negative placebo affect that compromises study outcomes [35, 141, 243]. In such a case, a preference-based study design may lead to better study outcomes due to improved adherence, and could be considered more individually ethical and equitable. However, abiding by patient preferences can also act as a confounder between the received treatment and the outcome, complicating statistical analysis.

3.3.2 Patient Population Heterogeneity and Trial Generalization. Participants of a clinical trial are often heterogeneous, differing from each other as well as from the outside population. Patients suffering from the same general conditions may have diverse health characteristics and levels of health risk [120, 133]. The prognoses of many diseases are highly variable. For example, COVID-19 ranges from non-symptomatic response to severe clinical courses requiring intensive care, sometimes leading to death [288]. Heterogeneity may also be the result of several other factors, including biased subject selection practices (Section 3.1) or differing patient risk preferences (Section 3.3.1). Patients who agree to enter clinical trials may be systematically distinct from those who refuse [163, 183, 243]. Additionally, patient population drift, where the characteristics of the trial participant population change over time, may arise with continuous patient recruitment [222], or trial expansion to different participant populations [259].

Heterogeneity may impact the internal validity of a trial; if the treatment effect varies across subgroups, its estimate is only meaningful for a well-defined population, and can be misleading for differing individuals [232]. Consequently, a minority of high-risk patients may have the majority of impact on trial outcomes [120, 121]. Heterogeneity of health risk may cause some patients to be unknowingly and unfairly susceptible to greater adverse effects of treatment [132, 133]. In an adaptive design, patients who enroll at different times may experience changed trial characteristics; participants may also be subject to trial modifications based upon preceding patients with different health characteristics [153].

Heterogeneity may also lead to problems with the external validity of a trial, such that the trial results may not apply to the general population, due to the specific characteristics of the trial participants [120, 133, 210, 266]. This negatively impacts not only the scientific validity of a trial by providing misleading results, but can also propagate health disparities as the impact of treatments on certain subgroups remain underexplored [46, 59, 68, 159].

Case 6: Cardiovascular disease trials and Medicare patients. In the US, Medicare is a federal program that provides health insurance coverage for the elderly. Cardiovascular disease is the leading cause of death in Medicare beneficiaries. The federal agency responsible for Medicare purports to make coverage decisions (funding of procedures and treatments) based upon evidence of improvement in health outcomes [68], typically derived from CTs. Researchers found that the participants of cardiovascular/heart failure CTs that inform Medicare coverage decisions differ significantly from the population of Medicare beneficiaries [68, 167]. Women and the elderly were significantly under-represented, with CT participants likely to be younger and male (avg. age 60.1, 75.4% male) than Medicare beneficiaries (avg. age 74.7, 48.1% male) [68]. Thus, decisions regarding the provision of treatments to patients (particularly elderly women) have been unfairly made on the basis of evidence that does not reflect the patient population [68].

Trial design and analysis methods may mitigate the impact of patient heterogeneity, although further research and adoption is crucial [132]. Subgroup analyses explore whether treatment effects differ based upon certain patient characteristics [91, 195]. In conventional subgroup analyses, participants are stratified using single patient characteristics, such as age; however, this type of analysis

has been criticized as inadequately informative [133, 210]. Risk-stratified subgroup analysis, where risk levels are determined by patient attributes, may be more representative [108, 121, 133, 211]. Unfortunately, a recent study has found a significant decrease in the usage of appropriate methods for subgroup analyses in RCTs over time [91]. Prognostic or risk factors may also be used in alternative randomization schemes, such as stratified randomization, where patients are divided into subgroups and randomized to subgroup-specific treatment arms [135], or novel adaptive enrichment designs, where recruitment is focuses on subgroups most likely to benefit [86, 186, 187].

Temporal heterogeneity particularly impacts trials that span a long duration, such as studies of rare diseases, which have embraced the use of RAR [4]. In RAR designs, unobserved time trends can inflate type-I error [226], even if restricted randomization is used [208]; temporal changes in trial data must therefore be carefully evaluated [257]. Even robust procedures [226] may be a computational burden and suffer form reductions in statistical power, indicating the need for computationally feasible testing procedures [257], or adaptive strategies that may incorporate non-stationarity [206].

4 OPPORTUNITIES FOR MACHINE LEARNING

In this section, we explore opportunities for ML to mitigate sources of unfairness (Section 3), emphasizing the consideration of fairness and ethics during development of ML. For each topic, we describe the problem setting, the current literature, highlight gaps in knowledge, and provide recommendations for future directions.

4.1 Subject Selection

Unfairness may arise throughout the process of subject selection, recruitment, and enrollment for clinical trials (Section 3.1). A large portion of ML research for clinical trials has focused on two problem settings: i) evaluation of eligibility criteria, and ii) patient identification for clinical research or trial matching. We consider these alongside the issues that may arise if fairness and bias are not carefully considered during ML development.

4.1.1 Evaluation of Eligibility Criteria. Categorical subject exclusion, typically codified in trial eligibility criteria, often induces unfairness in subject selection. Recent work has sought to use ML for data-driven evaluation and development of trial eligibility criteria to improve trial outcomes [138, 160]. Kim et al. [138] use EHR data to evaluate frequently used COVID-19 trial eligibility criteria, identifying alternative thresholds for eligibility criteria that would increase the number of desired observed outcomes with fewer patients. Liu et al. [160] use EHR data to evaluate the impact of eligibility criteria on cancer trial populations, finding they could broaden criteria without sacrificing trial efficacy. Such methods can broaden unnecessarily restrictive eligibility criteria and, in future work, potentially improve fairness by exposing whether certain sub-populations were made unjustifiably ineligible.

4.1.2 Patient Identification and Phenotyping. Data-driven ML methods may provide a path forward to mitigating under-representation in clinical trials through automation of subject selection. Researchers have used ML to identify patients eligible for clinical trials from EHR data, through matching of patient characteristics with trial criteria [105, 182, 286, 287] or existing trial participants [175]. Others have sought to identify patients who are most likely to agree to participate in a trial, for more efficient use of recruitment resources [181, 254]. Recently, researchers have expressed enthusiasm at the possibility of ML patient-matching algorithms to improve the diversity of trial cohorts [275].

Some research focuses on phenotyping algorithms, which aim to categorize patients based on health outcomes or disease states, as an intermediary to patient-trial matching. Practitioners have developed both supervised [157] and unsupervised [94] ML methods to learn disease phenotypes from EHR data, to identify representative cohorts that could benefit from proposed interventions. Phenotyping algorithms may also help in trials development by exposing subtypes of a disease to target for study. Li et al. [155] identify subtypes of Type 2 diabetes, a notably heterogeneous condition; this may inform treatment and research. Future work should focus upon developments to patient identification, phenotyping, and trial-matching algorithms to improve not only performance, but also fairness, explainability, and privacy.

4.1.3 Fair Practices. Exclusionary practices in clinical trials may bias resultant data that is used for downstream ML tasks; for example, researchers have sought to mitigate gender biases in word representations derived from trials data, which would be used for clinical prediction tasks [5]. EHR data also suffers from multiple sources of bias, including: i) *sampling bias*, where certain subgroups are over- or under-represented in the data, ii) *label bias*, where outcome variables have different meanings across subgroups, and iii) *feature bias*, where predictor variables have different meanings across subgroups [191]. Beyond its structural complexity, EHR data is also often plagued by missing data, where the absence of certain variables can lead to biases [253, 265]. Sampling bias may occur due to inequalities in access to care leading to less data collected from under-served populations [43, 111], or from research resulting from non-representative data cohorts [61]. Inequalities in the treatment of patients across subgroups may lead to label or feature bias, as outcomes and diagnostics may vary due to biases that impact quality of care [191, 271]. As racial/ethnic categories are social concepts that do not reliably inform genetic or clinical distinctions, they can be a source of feature bias; their use in clinical prediction algorithms may lead to negative impacts on marginalized populations [261].

ML practitioners must be careful when manually selecting target outcomes for healthcare research. Obermeyer et al. [185] found that a widely used algorithm to predict clinical risk for referral to care management programs displayed significant bias against Black patients. This occurred because the developers of the model used healthcare costs as a proxy for health outcome. However, in part due to entrenched societal disparities, less money is spent on Black patients with comparable health to white patients. Here, the choice of inappropriate labels led to biased predictions. This is relevant to the previously discussed works that use patient agreement of participation in a trial as an outcome label [181, 254]; predictions based upon this label may have unintended biases as, for example, a patient can only agree to participate if they are invited. Unfortunately, in comparison to White patients, a disproportionately small

proportion of racial/ethnic minorities are invited to participate in clinical trials, despite expressing a similar willingness to participate [268]. These concerns also apply to phenotyping algorithms that rely on disease labels, given biases present during diagnosis. As interest and adoption of ML for efficient subject selection grows, researchers must assess their methods for sources of potential bias on all fronts, including problem selection, data collection, and model development.

4.2 Eliciting and Incorporating Patient Risk Preferences

Below, we provide a review of the study of preference elicitation and describe how patient risk-benefit preferences (Section 3.3.1) may be incorporated into the design and development of clinical trials.

4.2.1 Risk Preference Elicitation. The study of preference elicitation–methods of collecting, determining, and modelling user preferences–varies across research fields. In the economics and psychology literature, preference elicitation is used to measure and evaluate the attitude of individuals regarding risk [41]. These methods can be categorized as *revealed preference methods*, which determine preferences based on behavioral data and *stated preference methods*, which rely on individuals to express their preferences using survey-based methods [8]. Revealed preferences may be considered the true preferences of individuals [8, 147], whereas stated preferences are commonly critiqued for their hypothetical nature. However, correspondence between the two can be high. In a study of vaccination preferences, Lambooij et al. [147] found that the stated preferences of 80% of individuals matched their revealed preferences.

In healthcare scenarios, individuals are usually asked to assess their risk attitudes towards treatments that have not yet occurred and that they have no prior experience with, precluding reliable revealed preference methods. Studies on patient health typically use stated preference methods to quantify patient *benefit-risk preferences* towards medical interventions [8, 106, 204, 228]. These methods can generally be categorized into two types: *direct-elicitation* methods, where participants are explicitly asked to identify acceptable levels of risk/benefit, and *conjoint-analysis* methods, where risk/benefit is assessed based on survey responses [106]. A direct-elicitation method is the *standard gamble* technique, wherein participants choose between a certain health state and a gamble that could result in either a better or worse outcome. The probabilities of the gamble are modified systematically until the participants are indifferent, allowing for estimation of the utility of the certain health state [106, 228]. A conjoint-analysis method is a *discrete-choice* experiment, wherein participants select their preferred scenario from a set of hypothetical choice situations [106, 228].

Researchers found that patients' reported preferences (versus true preferences) of treatment options can be affected by elicitation methodology [31], such as different framing of the same information [75]. Patients may also disagree on the associations between numerical values and verbal descriptions of risk, with one study reporting that different patients associated the phrase "frequent" within a range of 30−90% [75, 274]. Reliable quantification of patient risk preference is challenging; more research must be conducted to address the incongruence between reported and true preferences,

and to determine the underlying mechanisms behind individuals' responses to elicitation methods [113].

In the ML literature, preference elicitation focuses on models that represent user preferences [28, 45, 71, 101], often in the context of recommender systems, which learn preferences interactively. Attempts to automate the preference elicitation process across large populations include artificial neural networks with domain dependent priors [101], Bayesian preference elicitation by means of pairwise comparison queries [97], or supervised ML to improve direct elicitation results by recovering latent preferences [54]. These methods seek to correct stated preferences to more closely mimic unobserved revealed preferences. ML may also assist in combining stated preferences with clinical data (which may expose revealed preferences) to achieve better estimates of true preferences [106]; work has been pursued along these lines for non-health applications [8]. Ultimately, such software based solutions must also consider aspects of human-computer interaction. Stated preferences can be affected by interface design [196], suggesting that active user involvement in the interface design may yield more consistent and representative findings [37, 249]. In summary, computationally assisted preference elicitation is a delicate task that requires a multi-disciplinary perspective.

4.2.2 Incorporating Risk-Benefit Preferences. Elicited risk-benefit preferences may be used in different ways to support ethical treatment. Patient risk preferences are valuable in pre-trial clinical research; they provide information regarding treatment outcomes and clinical benefits that patients find most important, what side effects patients may be willing to cope with, and the types of treatments patients most prefer [170, 174, 231]. For example, a study of ovarian cancer patients found that patients preferred cure and extension of life over lessened symptoms from chemotherapy, exhibiting that patients prefer decreased mortality over decreased morbidity [174]. Stafinski et al. [231] compare patient and clinician views on the relative importance of treatment outcomes for cardiovascular conditions, finding differing risk-benefit thresholds. Sparano et al. [230] find that clinicians' reporting of symptomatic adverse events is incongruous with patient reported outcomes. Understanding patient risk profiles and how they relate to perceived clinical benefit is essential for more efficient and ethical choice of what treatments to pursue and compare within a clinical trial.

Trial designs that incorporate patient risk preferences into the treatment allocation mechanism have been proposed [35, 141, 243, 263]. In one design, patients that have strong preferences receive their treatment of choice, while others are randomized as usual [35, 141, 243]. In another method, patient risk preferences are elicited prior to treatment, and all participants are still randomized. This allows researchers to incorporate preference information into post-trial analysis [243], improving external validity without affecting internal validity [141, 263]. Novel preference-informed trials continue to be explored [7] and future work should also examine the ethics and statistical validity of preference-informed designs (given confounders) and how preferences may be incorporated into ML-driven adaptive designs.

Recent work in fairness in ML has sought to incorporate notions of preference into definitions of algorithmic fairness, resulting in classifier outcomes that meet certain preference guarantees [115, 140, 247, 283]. These works take inspiration from the economics notion of *envy-freeness*, asking that any group/individual prefers their treatment over that of others. Zafar et al. [283] and Ustun et al. [247] present fairness as a notion of envy-freeness among groups in the setting of decoupled classifiers, whereas Hossain et al. [115] impose fairness constraints using model outcomes. These works primarily deal with supervised ML models, where preference is defined as classifier performance metrics, such as accuracy. Kim et al. [140] incorporate individual outcome preferences, introducing a notion of preference-informed individual fairness that allows for deviations from similarity-based individual fairness, given alignment with individuals' preferences. Future work may address how preference-based fairness occurs in multi-armed bandit models, RL models, or in online learning, and incorporate specified risk preferences in online treatment allocation.

4.3 Modelling Adaptive Designs

The slow uptake of ADs, despite their benefits (Section 3.2.2), has been attributed to lack of familiarity, concerns of how funding bodies or regulators may view them, and lack of clarity regarding how ADs may be planned, implemented, interpreted, and reported in practice [36, 51, 187, 206]. This hesitancy is accompanied by a wariness against inclusion of ML in these designs, despite ongoing theoretical and methodological advancements [15, 19, 150, 206, 256, 272, 273].

Since the 1930s, ML models have been proposed for use in adaptive designs. Thompson sampling, also known as Bayesian response adaptive randomization, was first proposed as a solution to patient treatment allocation in clinical trials [240]. It is a solution strategy to the multi-armed bandit problem (MABP), where an agent trades off exploration (the acquisition of new information) with exploitation (optimal decisions based on existing knowledge). In a MABP, a fixed set of resources must be sequentially allocated between competing choices; in the trial setting, we may view the patients as the fixed set of resources and the trial treatment arms as the competing choices [256]. Although the MABP seems to encapsulate in theory the practical problems of trials, there has been continual hesitance to adopt them [10, 256]. The perceived shortcomings of bandit methods include loss of statistical power, challenges of statistical analysis (such as hypothesis testing) on bandit outcomes, and practical barriers to implementation (lack of access to statisticians or ML experts) [206, 256, 272].

Recent work has sought to innovate upon the field of multi-armed bandits (MABs) to support application to ADs, particularly RAR. Villar et al. [256] evaluate the performance of MABP approaches to RAR as compared to other allocation rules, including fixed randomization, finding that MABP methods improve patient welfare but reduce statistical power. The authors subsequently propose a solution building upon the Gittins index to improve statistical power [260]. Williamson et al. [273] also propose a bandit-based design that aims to improve statistical power and minimizes treatment effect biases. Williams et al. [272] expose the difficulties of

hypothesis testing on data derived from bandit models, and propose adjustments to existing statistical tests using knowledge of the bandit algorithm.

Theoretical ML research has focused mainly on applying MABs to dose-finding [15, 16, 19, 149, 150, 224] or treatment allocation [12, 197, 252]. From the viewpoint of multiple (sequential) testing, Xu et al. [276], Yang et al. [277] propose a unified MABP based framework that allows for online false discovery rate (FDR) control, i.e., a pre-specified FDR is guaranteed at any time of the sequential tests (even when stopping adaptively), yielding good sample complexity, high statistical power, robustness to various types of distribution shift, and low FDR at any point during the sequential testing. More broadly, recently renewed interest in anytime-valid sequential hypothesis testing (e.g., [146]) has surfaced practical methods for anytime-valid confidence intervals with FDR control even when adaptively combining or stopping sequences of statistical tests [96, 116, 128, 244]. Specifically for treatment effect estimation, Hadad et al. [99] present a method based on adaptive reweighting under adaptively collected data that aims at high statistical power and asymptotically correct coverage. While such methods can maintain clear statistical guarantees for arbitrary adaptive designs, we expect that bridging the gap between ML theory and clinical trials practice remains challenging due to a lack of relevant expertise from trial investigators, requirement of multi-disciplinary teams, and reluctance from regulators to consider the relatively new and theoretically involved underlying statistical concepts. Opportunities for ML in the development of ADs include improved sequential MAB or other treatment allocation methods that i) maintain required statistical guarantees under adaptive changes to the trial, ii) incorporate fairness considerations (addressing patient heterogeneity), and iii) handle or mitigate the impact of patient time trends (addressing patient drift).

4.4 Fair Exploration

In Section 4.3, we discussed the potential for explore-exploit methods (such as MABs or RL) to learn optimal policies by maximizing some notion of overall reward. Recent work has sought to characterize fairness in MABs [47, 93, 125, 162, 190] and RL [122]. These works evaluate fairness with respect to treatment arms/potential actions (rather than agents or users) from three general perspectives: i) *meritocratic fairness*, where a worse arm should never have a higher chance of being selected over a better one [125], ii) *individual fairness*, where similar individuals (in this case, arms) should be treated similarly [93, 162], and iii) constraints on the proportions in which arms should be pulled [47, 190]. However, the nature of the explore-exploit trade-off leads to unaddressed fairness implications for CTs, namely that certain subgroups may be subject to a greater *burden of exploration*, where they bear the negative consequences of exploration disproportionately [50]. The question of "how to explore and randomize ethically" is understudied from an ML perspective [50, 136]. Exploration often requires algorithms to take actions that may ultimately be sub-optimal in order to gather information. In CTs this may correspond to the sacrifice of one individual's well-being for the collective benefit of others. If we view it as a moral obligation to take only optimal actions for studied individuals, at the expense of exploration, this may slow learning

and lead to other types of unfairness due to a lack of information. Particularly in the clinical setting, exploration is at odds with the therapeutic obligation to treat patients as optimally as possible.

There has been comparatively sparse research on this conception of unfairness, which seeks to examine the burdens placed on certain subgroups, evaluating fairness from the perspective of the agents or users that are affected by the model outcomes, rather than the arms/actions. Jung et al. [126] demonstrate the problem of "free riding" in a multi-agent setting, where "free-riders" can incur minimal regret by accessing the information garnered by other agents. Raghavan et al. [199] introduce the concept of *group externalities* to quantify the negative impacts the presence of one group may impose on another, under the linear contextual bandits model. Baek and Farias [16] approach the problem in a similar fashion, introducing the concept of *grouped bandits*, where groupings are defined by user characteristics and access to different subsets of arms. The authors develop a solution to learn fair policies through use of the *Nash Social Welfare* (NSW) function, a notion of fairness in economics. In a slightly different approach, Hossain et al. [115] use the NSW to achieve fairness in the multi-agent MAB setting, where individuals/groups are considered separate agents with separate reward distributions. Future work should investigate the compatibility of these methods with real-world adaptive trial settings and the performance of these methods with respect to optimal treatment and statistical power.

4.5 Patient Heterogeneity and Study Generalization

Patient heterogeneity in CTs may lead to reduced internal and external validity, as well as fairness (Section 3.3.2). ML can provide methods of identifying heterogeneity pre- and post-trial, assessing generalizability, and aiding with generalization once heterogeneity is exposed. Patient heterogeneity, owing to a multitude of characteristics, results in *heterogeneity of treatment effects* (HTE), where treatment effects vary across individuals. In a clinical setting, the ultimate goal of HTE analysis is to estimate the causal effect for certain treatments on an individual level [134].

Kent et al. [134] identify two general approaches to predictive HTE analysis: i) *risk modelling*, where a multivariate model that predicts outcome risk is used to stratify the patient population [108, 119, 133], and ii) *effect modelling*, where treatment effects are predicted directly, sometimes incidentally identifying subgroups and predictive covariates [13, 236, 262]. While multivariate risk-stratified analysis is effective for identifying HTE [108, 133], it may suffer in the presence of other dimensions of risk, such as treatment-related harms, particularly if they are correlated with outcome risk [134]. Effect modelling can be better suited to account for heterogeneous risks or more individualized impacts on treatment effects [134]. Future work may focus on how ML can be practically incorporated into trial analyses; for example, Watson and Holmes [264] provide guidance on how ML can be used to determine HTE with strict type-I error control.

Other methods evaluate the generalizability of a clinical trial by assessing the representativeness of the trial population, comparing eligibility criteria with patients identified from large EHR datasets [156, 221]. Qi et al. [198] assess representativeness using

ML fairness metrics to identify under-represented subgroups. Such methods would benefit from further research on the interplay between representativeness, fairness, and generalizability.

Generalization of ML models to new populations/settings and dataset drift over time have been highlighted as broad challenges facing practical adoption of ML for healthcare [129]. Acquiring useful data for ML is complicated by noisy data collection processes and differences in formats of EHR and other clinical data [67], which could be standardized. Generalization may be improved through existing methods, including use of external validation datasets [63] or multiple data sites [178], and through identification and manipulation of covariates that impact generalizability [89]. ML may also assist in identification of dataset drift, and provide recommendations of appropriate model-updating procedures [62], which can then be resolved through other methods, such as periodic model validation and manual model re-training. Future research may focus on further automation of dataset drift correction or advancements in continual learning, i.e., study of ML systems that continually learn and evolve based on new data [90]. However, Futoma et al. [90] argue that overly generalized ML models may not be clinically useful. They assert that practitioners should focus on ensuring that the methodological process of model development is generalizable and that a ML model is well-understood—this would allow for practitioners to modify a model accordingly given a new setting or population. This holistic viewpoint may provide better clinical utility, helping practitioners use ML technology both correctly and carefully.

5 CONCLUSION

ML may improve the efficacy and fairness of CTs, given careful implementation and focused advancements in the field. Prior to more widespread adoption of ML for CTs, it is vital that researchers and practitioners consider both the possible benefits and possible harms the introduction of new technology can bring. The wider scientific community relies upon critical information derived from CTs to serve as ground truths for medical research. While we recognize the benefits of rapid discovery, CTs should be as representative and valid as possible, such that these ground truths are either applicable to all, not just a privileged minority, or at least come with explicit caveats. In the development of ML for clinical trials, ethics and fairness are critical considerations alongside any others such as performance, efficiency, or statistical validity, and should be recognized as key from problem selection to implementation in practice. For these reasons, we have provided a comprehensive overview of the sources of unfairness, as well as the accompanying opportunities for ML research that can be pursued, with fairness firmly in mind, to mitigate these issues. We caution that any such pursuits should be overseen by a multi-disciplinary team that can thoroughly evaluate the goals and consequences of such work.

REFERENCES

[1] [n. d.]. Clinical trials regulations. https://www.ukri.org/councils/mrc/guidance-for-applicants/policies-and-guidance-for-researchers/clinical-research-governance/clinical-trials-regulations/

[2] [n. d.]. Ethics in clinical research. https://clinicalcenter.nih.gov/recruit/ethics.html

[3] 2021. How clinical trials are approved. https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/how-clinical-trials-are-planned-and-organised/how-clinical-trials-are-approved

[4] Lusine Abrahamyan, Brian M Feldman, George Tomlinson, Marie E Faughnan, Sindhu R Johnson, Ivan R Diamond, and Samir Gupta. 2016. Alternative designs for clinical trials in rare diseases. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, Vol. 172. Wiley Online Library, 313–331.

[5] Shunit Agmon, Plia Gillis, Eric Horvitz, and Kira Radinsky. 2022. Gender-sensitive word embeddings for healthcare. *Journal of the American Medical Informatics Association : JAMIA* (2022).

[6] Al K Akobeng. 2005. Understanding randomised controlled trials. *Archives of disease in childhood* 90, 8 (2005), 840–844.

[7] Samina Ali, Gareth Hopkin, Naveen Poonai, Lawrence Richer, Maryna Yaskina, Anna Heath, Terry P Klassen, Christopher McCabe, Amy Gareth Jeff Martin Petros Eleanor David Marie-Chri Drendel Hopkin Round Offringa Pechlivanoglou Pulle, Amy L Drendel, Gareth Hopkin, Jeff Round, Martin Offringa, Petros Pechlivanoglou, Eleanor M. Pullenayegum, David Rios, Marie-Christine Auclair, Kelly Kim, Lise Bourrier, Lauren C. Dawson, Kamary Coriolano DaSilva, Pamela Marples, Rick Watts, Jennifer Thull-Freedman, Patrick J. McGrath, Timothy A. D. Graham, Lisa Hartling, Tannis Erickson, Brendon Foot, Kurt Schreiner, Julie Leung, Martin Petros Eleanor Juan David Jeff Offringa Pechlivanoglou Pullenayegum Rios Round, and Juan David Rios. 2021. A novel preference-informed complementary trial (PICT) design for clinical trial research influenced by strong patient preferences. *Trials* 22 (2021).

[8] Shehzad Ali and Sarah J Ronaldson. 2012. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *British medical bulletin* 103 1 (2012), 21–44.

[9] Derek C. Angus, Brian Michael Alexander, Scott M. Berry, Meredith Becker Buxton, Roger J. Lewis, Melissa Constance Paoloni, Steven A. R. Webb, Steven Arnold, Anna D. Barker, Donald A. Berry, Marc J. M. Bonten, Mary E. Brophy, Christopher Butler, Timothy F. Cloughesy, Lennie P. G. Derde, Laura J. Esserman, Ryan Ferguson, Louis D. Fiore, Sarah C. Gaffey, J. Michael Gaziano, Kathy E. Giusti, Herman Goossens, Stephane Heritier, Bradley Hyman, Michael Krams, Kay M. Larholt, Lisa M LaVange, Philip W. Lavori, Andrew W. Lo, Alex John London, Victoria G. Manax, Colin McArthur, Genevieve O'Neill, Giovanni Parmigiani, Jane Perlmutter, Elizabeth A. Petzold, Craig Ritchie, Kathryn M. Rowan, Christopher W. Seymour, Nathan I. Shapiro, Diane M Simeone, Bradley Smith, Bradley Spellberg, Ariel Dora Stern, Lorenzo Trippa, Mark R. Trusheim, Kert Viele, Patrick Yung Wen, and Janet Woodcock. 2019. Adaptive platform trials: definition, design, conduct and reporting considerations. *Nature Reviews Drug Discovery* 18 (2019), 797–807.

[10] Peter Armitage. 1985. The search for optimality in clinical trials. *International Statistical Review* 53 (1985), 15–24.

[11] World Medical Association et al. 1991. Declaration of helsinki. *Law, medicine & health care: a publication of the American Society of Law & Medicine* 19, 3-4 (1991), 264–265.

[12] Onur Atan, William R. Zame, and Mihaela van der Schaar. 2019. Sequential Patient Recruitment and Allocation for Adaptive Clinical Trials. In *AISTATS*.

[13] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113 (2016), 7353–7360.

[14] Andrew L Avins. 1998. Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials. *Journal of medical ethics* 24, 6 (1998), 401–408.

[15] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. 2021. On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials. *Journal of Machine Learning Research* 22 (2021), 1–38.

[16] Jackie Baek and V. Farias. 2021. Fair Exploration via Axiomatic Bargaining. *ArXiv* abs/2106.02553 (2021).

[17] Solon Barocas, Moritz Hardt, and A. Narayanan. 2018. Fairness and Machine Learning Limitations and Opportunities. In *Fairness and Machine Learning*.

[18] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

[19] Hamsa Bastani and Mohsen Bayati. 2015. Online Decision-Making with High-Dimensional Covariates. *Operations Research eJournal* (2015).

[20] Peter Bauer, Frank Bretz, Vladimir Dragalin, Franz König, and Gernot Wassmer. 2016. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine* 35, 3 (2016), 325–347.

[21] Françoise Baylis and Scott A Halperin. 2012. Research involving pregnant women: trials and tribulations. *Clinical investigation* 2 (2012), 139–146.

[22] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. 2020. Metric-free individual fairness in online learning. *arXiv preprint arXiv:2002.05474* (2020).

[23] John H. Beigel, Kay M Tomashek, Lori E. Dodd, Aneesh K. Mehta, Barry S. Zingman, Andre C. Kalil, Elizabeth L. Hohmann, Helen Y Chu, Anne F. Luetkemeyer, Susan E. Kline, Diego López de Castilla, Robert W. Finberg, Kerry L. Dierberg, Victor F. Tapson, Lanny Hsieh, Thomas F. Patterson, Roger Paredes, Daniel A. Sweeney, William R. Short, Giota Touloumi, David Chien Boon Lye, Norio Ohmagari, Myoung don Oh, Guillermo Ruiz-Palacios, Thomas Lars Benfield, Gerd Fätkenheuer, Mark G Kortepeter, Robert L. Atmar, Clarence Buddy Creech, Jens D. Lundgren, Abdel G. Babiker, Sarah Pett, James D. Neaton, Timothy H. Burgess, Tyler Bonnett, Michelle F. Green, Mat Makowski, Anu O. Osinusi, Seema U. Nayak, and Henry Clifford Lane. 2020. Remdesivir for the Treatment of Covid-19 — Final Report. *The New England Journal of Medicine* (2020).

[24] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366 (2019), 421 – 422.

[25] Deepak L Bhatt and Cyrus Mehta. 2016. Adaptive designs for clinical trials. *New England Journal of Medicine* 375, 1 (2016), 65–74.

[26] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. https://proceedings.mlr.press/v81/binns18a.html

[27] Atanu Biswas and Ayanendranath Basu. 2001. Robust adaptive designs in clinical trials for continuous responses. *Sankhyā: The Indian Journal of Statistics, Series B* (2001), 27–42.

[28] Avrim Blum, Jeffrey C. Jackson, Tuomas Sandholm, and Martin A. Zinkevich. 2003. Preference Elicitation and Query Learning. In *J. Mach. Learn. Res.*

[29] Danielle H Bodicoat, Ash C. Routen, Andrew Willis, Winifred Ekezie, Clare L Gillies, Claire Alexandra Lawson, Thomas Yates, Francesco Zaccardi, Melanie Jane Davies, and Kamlesh Khunti. 2021. Promoting inclusion in clinical trials—a rapid review of the literature and recommendations for action. *Trials* 22 (2021).

[30] Laura E. Bothwell and Aaron S. Kesselheim. 2017. The Real-World Ethics of Adaptive-Design Clinical Trials. *The Hastings Center report* 47 6 (2017), 27–37.

[31] Ann Bowling and Gene Rowe. 2005. "You decide doctor". What do patient preference arms in clinical trials really mean? *Journal of Epidemiology and Community Health* 59 (2005), 914 – 915.

[32] Valerie C Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* (2021), 1–6.

[33] Joseph V. Brady and Albert R. Jonsen. 2014. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists* 81 3 (2014), 4–13.

[34] Joel G. Breman, David L. Heymann, Graham Lloyd, Joseph B. McCormick, M Miatudila, Frederick A. Murphy, Jean-Jacques Muyembe-Tamfun, Peter Piot, J F Ruppol, Pierre Sureau, Guido van der Groen, and Karl M. Johnson. 2016. Discovery and Description of Ebola Zaire Virus in 1976 and Relevance to the West African Epidemic During 2013-2016. *The Journal of infectious diseases* 214 suppl 3 (2016), S93–S101.

[35] CR Brewin and C Bradley. 1989. Patient preferences and randomised clinical trials. *British Medical Journal* 299 (1989), 313 – 315.

[36] Thomas Burnett, Pavel Mozgunov, Philip Pallmann, Sofia S Villar, Graham M Wheeler, and Thomas Jaki. 2020. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC medicine* 18, 1 (2020), 1–21.

[37] Carrie J. Cai, Emily Reif, Narayan Hegde, J. Hipp, Been Kim, D. Smilkov, M. Wattenberg, F. Viégas, G. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[38] D Ross Camidge, Haeseong Park, Karen E Smoyer, Ira Jacobs, Lauren J Lee, Zemfira Askerova, Justin McGinnis, and Yousef Zakharia. 2021. Race and ethnicity representation in clinical trials: findings from a literature review of Phase I oncology trials. *Future Oncology* 0 (2021).

[39] An-Wen Chan, Jennifer Marie Tetzlaff, Douglas G. Altman, Andreas Laupacis, Peter Christian Gøtzsche, Karmela Krleza-Jeric, Asbjørn Hróbjartsson, Howard Mann, Kay Dickersin, Jesse A. Berlin, Caroline J. Doré, Wendy R. Parulekar, William S M Summerskill, Trish Groves, Kenneth F. Schulz, Harold C. Sox, Frank W. Rockhold, Drummond Rennie, and David Moher. 2013. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of internal medicine* 158 3 (2013), 200–7.

[40] Danton Char, N. Shah, and D. Magnus. 2018. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *The New England journal of*

medicine 378 11 (2018), 981–983.

[41] Gary Charness, Uri Gneezy, and Alex Imas. 2013. Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization* 87 (2013), 43–51.

[42] I. Chen, Fredrik D. Johansson, and D. Sontag. 2018. Why Is My Classifier Discriminatory?. In *NeurIPS*.

[43] I. Chen, E. Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and M. Ghassemi. 2021. Ethical Machine Learning in Health Care. *Annual review of biomedical data science* 4 (2021), 123–144.

[44] I. Chen, Peter Szolovits, and M. Ghassemi. 2019. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA journal of ethics* 21 2 (2019), E167–179.

[45] Li Chen and Pearl Pu. 2004. Survey of Preference Elicitation Methods.

[46] Moon S Chen, Primo N. Jr. Lara, Julie H T Dang, Debora A. Paterniti, and Karen Kelly. 2014. Twenty Years Post-NIH Revitalization Act: Renewing the Case for Enhancing Minority Participation in Cancer Clinical Trials. *Cancer* 120 (2014), 1091.

[47] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and S. Nikolaidis. 2020. Fair Contextual Multi-Armed Bandits: Theory and Experiments. In *UAI*.

[48] Leslie Cho, Amanda R Vest, Michelle L O'Donoghue, Modele O Ogunniyi, Amy A Sarma, Kara J Denby, Emily S Lau, Jeanne E Poole, Kathryn J Lindley, Roxana Mehran, et al. 2021. Increasing Participation of Women in Cardiovascular Trials: JACC Council Perspectives. *Journal of the American College of Cardiology* 78, 7 (2021), 737–751.

[49] A. Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5 2 (2017), 153–163.

[50] A. Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *ArXiv* abs/1810.08810 (2018).

[51] S. Chow and M. Chang. 2008. Adaptive design methods in clinical trials – a review. *Orphanet Journal of Rare Diseases* 3 (2008), 11 – 11.

[52] Nicholas A. Christakis and Morris J. Panner. 1991. Existing International Ethical Guidelines for Human Subjects Research: Some Open Questions. *Law, Medicine and Health Care* 19 (1991), 214 – 221.

[53] Luther T Clark, Laurence Watkins, Ileana L. Piña, Mary E Elmer, Olakunle O. Akinboboye, Millicent Gorham, Brenda D. Jamerson, Cassandra McCullough, Christine Pierre, Adam B. Polis, Gary A. Puckrein, and Jeanne M. Regnante. 2019. Increasing Diversity in Clinical Trials: Overcoming Critical Barriers. *Current problems in cardiology* 44 5 (2019), 148–172.

[54] John A. Clithero, Jae Joon Lee, and Joshua Tasoff. 2019. Supervised Machine Learning for Eliciting Individual Demand. *Computational Materials Science eJournal* (2019).

[55] Nuremberg Code. 1949. The Nuremberg Code. *Trials of war criminals before the Nuremberg military tribunals under control council law* 10, 1949 (1949), 181–2.

[56] Office of the Commissioner. [n. d.]. Step 3: Clinical research. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research

[57] Edward M Connor, Rhoda S Sperling, Richard Gelber, Pavel Kiselev, Gwendolyn Scott, Mary Jo O'sullivan, Russell VanDyke, Mohammed Bey, William Shearer, Robert L Jacobson, et al. 1994. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *New England Journal of Medicine* 331, 18 (1994), 1173–1180.

[58] S. Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv* abs/1808.00023 (2018).

[59] Paul Cotton. 1990. Examples abound of gaps in medical knowledge because of groups excluded from scientific study. *JAMA* 263, 8 (1990), 1051–1055.

[60] Ricardo Eccard da Silva, Angélica Amorim Amato, Dirce Bellezi Guilhem, and Maria Rita Carvalho Garbi Novaes. 2016. Globalization of clinical trials: ethical and regulatory implications. *International Journal of Clinical Trials* 3 (2016), 1–8.

[61] Ralph B. D'Agostino, Scott M. Grundy, Lisa M Sullivan, and Pwf Wilson. 2001. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 286 2 (2001), 180–7.

[62] Sharon E. Davis, Robert A. Greevy, Christopher J Fonnesbeck, Thomas A. Lasko, Colin G. Walsh, and Michael E. Matheny. 2019. A nonparametric updating method to correct clinical prediction model drift. *Journal of the American Medical Informatics Association : JAMIA* (2019).

[63] Thomas P. A. Debray, Yvonne Vergouwe, Hendrik Koffijberg, Daan Nieboer, Ewout Willem Steyerberg, and Karel G. M. Moons. 2015. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of clinical epidemiology* 68 3 (2015), 279–89.

[64] Nina Deliu, Joseph J Williams, and Sofia S Villar. 2021. Efficient Inference Without Trading-off Regret in Bandits: An Allocation Probability Test for Thompson Sampling. *arXiv preprint arXiv:2111.00137* (2021).

[65] Colleen C Denny and Christine Grady. 2007. Clinical research with economically disadvantaged populations. *Journal of medical ethics* 33, 7 (2007), 382–385.

[66] C. DeSantis, R. Siegel, A. Sauer, K. Miller, S. Fedewa, Kassandra I. Alcaraz, and A. Jemal. 2016. Cancer statistics for African Americans, 2016: Progress and

opportunities in reducing racial disparities. *CA: A Cancer Journal for Clinicians* 66 (2016).

[67] Gregory P. Dexter, Shaun J. Grannis, Brian E. Dixon, and Suranga Nath Kasthurirathne. 2020. Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2020 (2020), 152–161.

[68] Sanket S. Dhruva and Rita F. Redberg. 2008. Variations between clinical trial participants and Medicare beneficiaries in evidence used for Medicare national coverage decisions. *Archives of internal medicine* 168 2 (2008), 136–40.

[69] M Diaz and D Neuhauser. 2005. Pasteur and parachutes: when statistical process control is better than a randomized controlled trial. *BMJ Quality & Safety* 14, 2 (2005), 140–143.

[70] Eric L. Ding, Neil R. Powe, JoAnn E. Manson, Noëlle S. Sherber, and Joel B. Braunstein. 2007. Sex differences in perceived risks, distrust, and willingness to participate in clinical trials: a randomized study of cardiovascular prevention trials. *Archives of internal medicine* 167 9 (2007), 905–12.

[71] Paolo Dragone, Stefano Teso, and Andrea Passerini. 2018. Constructive Preference Elicitation. *Frontiers Robotics AI* 4 (2018), 71.

[72] Rebecca Dresser. 1992. Wanted single, white male for medical research. *The Hastings Center Report* 22, 1 (1992), 24–29.

[73] Narjust Duma, Jesus Vera Aguilera, Jonas Paludo, Candace L Haddox, Miguel Gonzalez Velez, Yucai Wang, Konstantinos Leventakos, Joleen M Hubbard, Aaron S. Mansfield, Ronald S. Go, and Alex A. Adjei. 2018. Representation of Minorities and Women in Oncology Clinical Trials: Review of the Past 14 Years. *Journal of oncology practice* 14 1 (2018), e1–e10.

[74] C. Dwork, Moritz Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. Fairness through awareness. *ArXiv* abs/1104.3913 (2012).

[75] Adrian G. K. Edwards and Glyn Elwyn. 2001. Understanding risk and lessons for clinical risk communication about treatment preferences. *BMJ Quality & Safety* 10 (2001).

[76] Ezekiel J. Emanuel, David Wendler, and Christine Grady. 2000. What makes clinical research ethical? *JAMA* 283 20 (2000), 2701–11.

[77] Steven Epstein. 2008. The Rise of 'Recruitmentology': Clinical Research, Racial Knowledge, and the Politics of Inclusion and Difference. *Social Studies of Science* 38, 5 (2008), 801–832. https://doi.org/10.1177/0306312708091930 arXiv:https://doi.org/10.1177/0306312708091930 PMID: 19227822.

[78] Azure N. Faucette, Michael D Pawlitz, B. Pei, Fayi Yao, and Kang Chen. 2015. Immunization of pregnant women: Future of early infant protection. *Human Vaccines & Immunotherapeutics* 11 (2015), 2549 – 2555.

[79] Azure N. Faucette, Benjamin L. Unger, Bernard Gonik, and Kang Chen. 2015. Maternal vaccination: moving the science forward. *Human reproduction update* 21 1 (2015), 119–35.

[80] Michael Feldman, Sorelle A. Friedler, John Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015).

[81] Kadija Ferryman. 2018. Artificial intelligence can entrench disparities—here's what we must do. *The Cancer Letter* 44 (2018), 543.

[82] Jill A. Fisher and Corey A. Kalbaugh. 2011. Challenging assumptions about minority participation in US clinical research. *American journal of public health* 101 12 (2011), 2217–22.

[83] Center for Drug Evaluation and Research. [n. d.]. (irbs) and protection of human subjects. https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials

[84] Council for International Organizations of Medical Sciences. 2016. *International Ethical Guidelines for Health-related Research Involving Humans*. Technical Report. Council for International Organizations of Medical Sciences, Geneva, Switzerland. http://www.cioms.ch/ethical-guidelines-2016/

[85] Benjamin Freedman. 1987. Equipoise and the ethics of clinical research. *The New England journal of medicine* 317 3 (1987), 141–5.

[86] Tim Friede, Nigel Stallard, and Nicholas Parsons. 2020. Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. *Biometrical Journal* 62, 5 (2020), 1264–1283.

[87] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[88] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[89] Joseph D. Futoma, Morgan G. Simons, Finale Doshi-Velez, and Rishikesan Kamaleswaran. 2021. Generalization in Clinical Prediction Models: The Blessing and Curse of Measurement Indicator Variables. *Critical Care Explorations* 3 (2021).

[90] Joseph D. Futoma, Morgan G. Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet. Digital Health* 2 (2020), e489 –

e492.

[91] Nicole B Gabler, Naihua Duan, Eli Raneses, Leah H Suttner, Michael Ciarametaro, Elizabeth Cooney, Robert W. Dubois, Scott D. Halpern, and Richard L. Kravitz. 2016. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. *Trials* 17 (2016).

[92] Judy Wawira Gichoya, Liam G. McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. 2021. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ health & care informatics* 28 1 (2021).

[93] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online Learning with an Unknown Fairness Metric. *ArXiv* abs/1802.06936 (2018).

[94] Benjamin Scott Glicksberg, Riccardo Miotto, Kipp W. Johnson, Khader Shameer, Li Li, Rongbing Chen, and Joel T. Dudley. 2018. Automated disease cohort selection using word embeddings from Electronic Health Records. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23 (2018), 145–156.

[95] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[96] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. 2020. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*. IEEE, 1–54.

[97] Shengbo Guo and Scott Sanner. 2010. Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries. In *AISTATS*.

[98] Marya Viorst Gwadz, Keith Cylar, Noelle Regina Leonard, Marion Riedel, Nina Herzog, Gricel N. Arredondo, Charles M. Cleland, M Aguirre, A M Marshak, Donna Mildvan, and The Heart to Heart Collaborative Research Team. 2009. An Exploratory Behavioral Intervention Trial to Improve Rates of Screening for AIDS Clinical Trials Among Racial/Ethnic Minority and Female Persons Living with HIV/AIDS. *AIDS and Behavior* 14 (2009), 639–648.

[99] Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. 2021. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences of the United States of America* 118 (2021).

[100] Tufia C. Haddad, Jane M Helgeson, Katharine E Pomerleau, Marissa Makey, Phillip Lombardo, Sadie Coverdill, Alexandra Urman, Melissa Rammage, Matthew P. Goetz, and Nicholas F. LaRusso. 2018. Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice. *Journal of Clinical Oncology* 36 (2018), 6550–6550.

[101] Peter Haddawy, Vu A. Ha, Angelo C. Restificar, Benjamin Geisler, and John M. Miyamoto. 2003. Preference Elicitation via Theory Refinement. *J. Mach. Learn. Res.* 4 (2003), 317–337.

[102] Lauren M. Hamel, Louis A Penner, Terrance L Albrecht, Elisabeth I. Heath, Clement K. Gwede, and Susan Eggly. 2016. Barriers to Clinical Trial Enrollment in Racial and Ethnic Minority Patients With Cancer. *Cancer control : journal of the Moffitt Cancer Center* 23 4 (2016), 327–337.

[103] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

[104] Romana Hasnain-Wynia, David W. Baker, David R. Nerenz, Joe M Feinglass, Anne C. Beal, Mary Beth Landrum, Raj Behal, and Joel S. Weissman. 2007. Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance measures. *Archives of internal medicine* 167 12 (2007), 1233–9.

[105] Hamed Hassanzadeh, Sarvnaz Karimi, and Anthony N. Nguyen. 2020. Matching patients to clinical trials using semantically enriched document representation. *Journal of biomedical informatics* (2020), 103406.

[106] A. Brett Hauber, Angelyn O. Fairchild, and F. Reed Johnson. 2013. Quantifying Benefit–Risk Preferences for Medical Interventions: An Overview of a Growing Empirical Literature. *Applied Health Economics and Health Policy* 11 (2013), 319–329.

[107] A. Brett Hauber, F. Reed Johnson, Howard M Fillit, Ateesha Farah Mohamed, Christopher W. Leibman, H. Michael Arrighi, Michael Grundman, and Raymond J Townsend. 2009. Older Americans' Risk-benefit Preferences for Modifying the Course of Alzheimer Disease. *Alzheimer Disease & Associated Disorders* 23 (2009), 23–32.

[108] Rodney A. Hayward, David M. Kent, Sandeep Vijan, and Timothy P. Hofer. 2006. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology* 6 (2006), 18 – 18.

[109] Asefeh Heiat, Cary P. Gross, and Harlan M. Krumholz. 2002. Representation of the elderly, women, and minorities in heart failure clinical trials. *Archives of internal medicine* 162 15 (2002), 1682–8.

[110] Gail E. Henderson, Larry R. Churchill, Arlene M. Davis, Michele M Easter, Christine Grady, Steven Joffe, Nancy E. Kass, Nancy M P King, Charles W. Lidz, Franklin G. Miller, Daniel K. Nelson, Jeffrey Peppercorn, Barbra Bluestone Rothschild, Pamela L Sankar, Benjamin S. Wilfond, and Catherine R. Zimmer. 2007. Clinical Trials and Medical Care: Defining the Therapeutic Misconception. *PLoS Medicine* 4 (2007).

[111] E. Hing and E. Burt. 2009. Are There Patient Disparities When Electronic Health Records Are Adopted? *Journal of Health Care for the Poor and Underserved* 20

[112] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[113] Felix Holzmeister and Matthias Stefan. 2019. The Risk Elicitation Puzzle Revisited: Across-Methods (In)consistency?

[114] Monica Webb Hooper, Taghrid Asfar, Marina Unrod, Asha Dorsey, John B. Correa, Karen O Brandon, Vani N. Simmons, Michael A Antoni, Tulay Koru-Sengul, David J Lee, and Thomas H Brandon. 2019. Reasons for Exclusion from a Smoking Cessation Trial: An Analysis by Race/Ethnicity. *Ethnicity & disease* 29 1 (2019), 23–30.

[115] Safwan Hossain, Andjela Mladenovic, and Nisarg Shah. 2020. Designing Fairly Fair Classifiers Via Economic Fairness Notions. *Proceedings of The Web Conference 2020* (2020).

[116] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2021. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* 49, 2 (2021), 1055–1080.

[117] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770* (2020).

[118] Christina Ilvento. 2019. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250* (2019).

[119] Marjorie Z Imperial, Patrick P. J. Phillips, Payam Nahid, and Radojka M. Savic. 2021. Precision-Enhancing Risk Stratification Tools for Selecting Optimal Treatment Durations in Tuberculosis Clinical Trials. *American Journal of Respiratory and Critical Care Medicine* 204 (2021), 1086 – 1096.

[120] Jpa Ioannidis and Joseph Lau. 1997. The impact of high-risk patients on the results of clinical trials. *Journal of clinical epidemiology* 50 10 (1997), 1089–98.

[121] John P. A. Ioannidis and Joseph Lau. 1998. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *American journal of epidemiology* 148 11 (1998), 1117–26.

[122] S. Jabbari, Matthew Joseph, M. Kearns, Jamie H. Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *ICML*.

[123] Maud O Jansen, Peter Angelos, Stephen J Schrantz, Jessica S Donington, Maria Lucia L Madariaga, and Tanya L Zakrison. 2021. Fair and equitable subject selection in concurrent COVID-19 clinical trials. *Journal of Medical Ethics* 47, 1 (2021), 7–11.

[124] F. Reed Johnson, George Van Houtven, Semra Özdemir, Steven L. Hass, Jeffrey White, Gordon S. Francis, David W. Miller, and J. Theodore Phillips. 2009. Multiple sclerosis patients—benefit-risk preferences: Serious adverse event risks versus treatment efficacy. *Journal of Neurology* 256 (2009), 554–562.

[125] Matthew Joseph, M. Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. *ArXiv* abs/1605.07139 (2016).

[126] Christopher Jung, Sampath Kannan, and Neil Lutz. 2020. Quantifying the Burden of Exploration and the Unfairness of Free Riding. *ArXiv* abs/1810.08743 (2020).

[127] Leslie A Kalish and Colin B Begg. 1985. Treatment allocation methods in clinical trials: a review. *Statistics in Medicine* 4, 2 (1985), 129–144.

[128] Eugene Katsevich and Aaditya Ramdas. 2020. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics* 48, 6 (2020), 3465–3487.

[129] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 17 (2019).

[130] J M Kendall. 2003. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal* 20 (2003), 164 – 168.

[131] Tessa Kennedy-Martin, Sarah E. Curtis, Douglas E Faries, Susan Robinson, and Joseph A Johnston. 2015. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 16 (2015).

[132] David M. Kent, Alawi A. Alsheikh-Ali, and Rodney A. Hayward. 2008. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* 9 (2008), 30 – 30.

[133] David M. Kent and Rodney A. Hayward. 2007. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 298 10 (2007), 1209–12.

[134] David M. Kent, Ewout Willem Steyerberg, and David van Klaveren. 2018. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *British Medical Journal* 363 (2018).

[135] Walter N. Kernan, Catherine M. Viscoli, Robert W. Makuch, Lawrence M. Brass, and Ralph I Horwitz. 1999. Stratified randomization for clinical trials. *Journal of clinical epidemiology* 52 1 (1999), 19–26.

[136] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 277–287.

[137] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*,

(2009), 473 – 488.

I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf

[138] Jae Hyun Kim, Casey N. Ta, Cong Liu, Cynthia Sung, Alex M. Butler, Latoya A. Stewart, Lyudmila Ena, James R. Rogers, Junghwan Lee, Anna Ostropolets, Patrick B. Ryan, Hao Liu, Shing M. Lee, Mitchell S. V. Elkind, and Chunhua Weng. 2021. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. *Journal of the American Medical Informatics Association : JAMIA* 28 (2021), 14 – 22.

[139] Michael P. Kim, Amirata Ghorbani, and James Y. Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019).

[140] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and G. Yona. 2020. Preference-Informed Fairness. In *ITCS*.

[141] Michael King, Irwin Nazareth, Fiona Lampe, Peter Bower, Martin Chandler, M Morou, Bonnie Sibbald, and Rosalind Lai. 2005. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *JAMA* 293 9 (2005), 1089–99.

[142] J. Kleinberg, S. Mullainathan, and M. Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv* abs/1609.05807 (2017).

[143] Stanley G Korenman. 2006. Teaching the responsible conduct of research in humans (RCRH). *Los Angeles: University of California* (2006).

[144] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*.

[145] Thomas Laage, John W Loewy, Sandeep Menon, Eva R Miller, Erik Pulkstenis, Natalia Kan-Dobrosky, and Christopher Coffey. 2017. Ethical considerations in adaptive design clinical trials. *Therapeutic innovation & regulatory science* 51, 2 (2017), 190–199.

[146] T.L Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6, 1 (1985), 4–22. https://doi.org/10.1016/0196-8858(85)90002-8

[147] Mattijs S. Lambooij, Irene Anhai Harmsen, Jorien Veldwijk, Hester E. de Melker, Liesbeth Mollema, Yolanda WM van Weert, and G. Ardine de Wit. 2015. Consistency between stated and revealed preferences: a discrete choice experiment and a behavioural experiment on vaccination behaviour compared. *BMC Medical Research Methodology* 15 (2015).

[148] Hyun-Suk Lee, Cong Shen, James Jordon, and Mihaela Schaar. 2020. Contextual constrained learning for dose-finding clinical trials. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2645–2654.

[149] Hyun-Suk Lee, Cong Shen, James Jordon, and Mihaela van der Schaar. 2020. Contextual Constrained Learning for Dose-Finding Clinical Trials. In *AISTATS*.

[150] Hyun-Suk Lee, Cong Shen, William R. Zame, Jang-Won Lee, and Mihaela van der Schaar. 2021. SDF-Bayes: Cautious Optimism in Safe Dose-Finding Clinical Trials with Drug Combinations and Heterogeneous Patient Groups. In *AISTATS*.

[151] Jinhee Lee, Hanwul Shin, Jun Young Lee, Jae Seok Kim, Jae-Won Yang, Keumwha Lee, Andreas Kronbichler, and Jae Il Shin. 2020. A Comprehensive Analysis of Clinical Trials in the COVID-19 Pandemic Era. *Medicina* 56 (2020).

[152] J Jack Lee. 2015. Commentary on Hey and Kimmelman. *Clinical Trials* 12, 2 (2015), 110–112.

[153] Laurie J. Legocki, William J. Meurer, Shirley M. Frederiksen, Roger J. Lewis, Valerie L. Durkalski, Donald A. Berry, William G. Barsan, and Michael Derwin Fetters. 2015. Clinical trialist perspectives on the ethics of adaptive clinical trials: a mixed-methods analysis. *BMC Medical Ethics* 16 (2015).

[154] Robert J. Levine. 2002. International ethical guidelines for biomedical research involving human subjects. *Bulletin of medical ethics* 182 (2002), 17–23.

[155] Lei Li, Wei-Yi Cheng, Benjamin S. Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P. Bottinger, and Joel T. Dudley. 2015. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine* 7 (2015), 311ra174 – 311ra174.

[156] Qian Li, Zhe He, Yi Guo, Hansi Zhang, Thomas J. George, William R. Hogan, Neil Charness, and J. Bian. 2019. Assessing the Validity of a a priori Patient-Trial Generalizability Score using Real-world Data from a Large Clinical Data Research Network: A Colorectal Cancer Clinical Trial Case Study. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2019 (2019), 1101–1110.

[157] Katherine P. Liao, Tianxi Cai, Guergana K. Savova, Shawn N. Murphy, Elizabeth W. Karlson, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Stanley Y. Shaw, Zongqi Xia, Peter Szolovits, Susanne E. Churchill, and Isaac S. Kohane. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *The BMJ* 350 (2015).

[158] Richard J Lilford and Jennifer Jackson. 1995. Equipoise and the ethics of randomization. *Journal of the Royal Society of Medicine* 88, 10 (1995), 552.

[159] Abby Lippman. 2006. The Inclusion of Women in Clinical Trials: Are We Asking the Right Questions?

[160] Ruishan Liu, Shemra Rizzo, Sam Whipple, Navdeep Pal, Arturo López Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William B. Capra, Ryan Copping, and James Zou. 2021. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* (2021).

[161] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie Jane Calvert, Alastair K. Denniston, Hutan Ashrafian, Andrew Beam, An-Wen Chan, Gary S. Collins, Ara Darzi, Jonathan J. Deeks, Mohammed K. Elzarrad, Cyrus Espinoza, Andre Esteva, Livia Faes, Lavinia Ferrante di Ruffano, J Fletcher, Robert M. Golub, Hugh Harvey, Charlotte J Haug, Christopher Holmes, Adrian Jonas, Pearse A. Keane, Christopher J. Kelly, Aaron Y. Lee, Cecilia S. Lee, Elaine Manna, James Matcham, Melissa D. McCradden, Joao Monteiro, Cynthia D. Mulrow, Luke Oakden-Rayner, Dina N. Paltoo, Maria Beatrice Panico, Gary Price, Samuel d. Rowley, Richard Savage, Rupa Sarkar, Sebastian J. Vollmer, and Christopher Yau. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *The BMJ* 370 (2020).

[162] Y. Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and D. Parkes. 2017. Calibrated Fairness in Bandits. *ArXiv* abs/1707.01875 (2017).

[163] Hilary Llewellyn-Thomas, M J McGreal, Elaine C. Thiel, Sheldon Fine, and Charles Erlichman. 1991. Patients' willingness to enter clinical trials: measuring the association with perceived benefit and preference for decision participation. *Social science & medicine* 32 1 (1991), 35–42.

[164] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal Reasoning for Algorithmic Fairness. *ArXiv* abs/1805.05859 (2018).

[165] Alex John London. 2018. Learning health systems, clinical equipoise and the ethics of response adaptive randomisation. *Journal of medical ethics* 44, 6 (2018), 409–415.

[166] David Madras, Elliot Creager, T. Pitassi, and R. Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*.

[167] Frederick A. Masoudi, Edward P. Havranek, Pam Wolfe, Cary P. Gross, Saif S. Rathore, John F. Steiner, Diana L. Ordin, and Harlan M. Krumholz. 2003. Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure. *American heart journal* 146 2 (2003), 250–7.

[168] Charles R. Mccarthy. 1994. Historical background of clinical trials involving women and minorities. *Academic Medicine* 69 (1994), 695–8.

[169] Gillian McKay and Shelley Lees. 2020. Protection by Exclusion: Another Missed Opportunity to Include Pregnant Women in Research during the Coronavirus Disease 2019 (COVID-19) Pandemic. *Obstetrics & Gynecology* (2020).

[170] Richard P. McQuellon, Hyman B. Muss, S. Hoffman, Greg Russell, B. Cathy Craven, and Suzanne B. Yellen. 1995. Patient preferences for treatment of metastatic breast cancer: a study of women with early-stage breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 13 4 (1995), 858–68.

[171] Ninareh Mehrabi, Fred Morstatter, N. Saxena, Kristina Lerman, and A. Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54 (2021), 1 – 35.

[172] William J. Meurer, Roger J. Lewis, and Donald A. Berry. 2012. Adaptive clinical trials: a partial remedy for the therapeutic misconception? *JAMA* 307 22 (2012), 2377–8.

[173] Franklin G. Miller and Howard Brody. 2003. A critique of clinical equipoise. Therapeutic misconception in the ethics of clinical trials. *The Hastings Center report* 33 3 (2003), 19–28.

[174] Lindsey E. Minion, Robert L. Coleman, Ronald D. Alvarez, and Thomas J. Herzog. 2016. Endpoints in clinical trials: What do patients consider important? A survey of the Ovarian Cancer National Alliance. *Gynecologic oncology* 140 2 (2016), 193–8.

[175] Riccardo Miotto and Chunhua Weng. 2015. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *Journal of the American Medical Informatics Association : JAMIA* 22 (2015), e141 – e150.

[176] Jessica Mozersky, Douglas L Mann, and James M DuBois. 2020. The National Institute of Allergy and Infectious Diseases decision to stop the adaptive COVID-19 trial: on solid ethical and scientific grounds. , 645–647 pages.

[177] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*. PMLR, 7097–7107.

[178] Takashi Nakano, Masahiro Takamura, Naho Ichikawa, Go Okada, Yasumasa Okamoto, Makiko Yamada, Tetsuya Suhara, Shigeto Yamawaki, and Junichiro Yoshimoto. 2020. Enhancing Multi-Center Generalization of Machine Learning-Based Depression Diagnosis From Resting-State fMRI. *Frontiers in Psychiatry* 11 (2020).

[179] Cecilia Nardini. 2014. The ethics of clinical trials. *ecancermedicalscience* 8 (2014).

[180] Alan R. Nelson. 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association* 94 8 (2002), 666–8.

[181] Yizhao Ni, Andrew F. Beck, Regina Taylor, Jenna Dyas, Imre Solti, and Judith W. Dexheimer. 2015. Will they participate? Predicting patients' response to clinical trial invitations in a pediatric emergency department. *Journal of the American Medical Informatics Association : JAMIA* 23 (2015), 671 – 680.

[182] Yizhao Ni, Stephanie Spellman Kennebeck, Judith W. Dexheimer, Constance M. McAneney, Huaxiu Tang, Todd Lingren, Qi Li, Haijun Zhai, and Imre Solti. 2015. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. *Journal of*

the American Medical Informatics Association : JAMIA 22 (2015), 166 – 178.

[183] Ryan D. Nipp, Kessely Hong, and Electra D. Paskett. 2019. Overcoming Barriers to Clinical Trial Enrollment. American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting 39 (2019), 105–114.

[184] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine 375, 13 (2016), 1216.

[185] Z. Obermeyer, Brian W. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (2019), 447 – 453.

[186] Thomas Ondra, Alex Dmitrienko, Tim Friede, Alexandra Graf, Frank Miller, Nigel Stallard, and Martin Posch. 2016. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. Journal of biopharmaceutical statistics 26, 1 (2016), 99–119.

[187] Philip Pallmann, Alun W Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V Hampson, Jane Holmes, Adrian P Mander, Lang'o Odondi, Matthew R Sydes, et al. 2018. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC medicine 16, 1 (2018), 1–15.

[188] Asger Paludan-Müller, David Ruben Teindl Laursen, and Asbjørn Hróbjartsson. 2016. Mechanisms and direction of allocation bias in randomised clinical trials. BMC medical research methodology 16, 1 (2016), 1–10.

[189] Jay J H Park, Kristian Thorlund, and Edward J. Mills. 2018. Critical concepts in adaptive clinical trials. Clinical Epidemiology 10 (2018), 343 – 351.

[190] Vishakha Patil, Ganesh Ghalme, Vineet J. Nair, and Y. Narahari. 2020. Achieving Fairness in the Stochastic Multi-armed Bandit Problem. ArXiv abs/1907.10516 (2020).

[191] J. Paulus and D. Kent. 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digital Medicine 3 (2020).

[192] J. Pearl. 2000. Causality: Models, Reasoning and Inference. In None.

[193] Emily Peckham, Sally Brabyn, Liz Cook, Thomas Devlin, Jo Dumville, and David J Torgerson. 2015. The use of unequal randomisation in clinical trials—an update. Contemporary clinical trials 45 (2015), 113–122.

[194] S. Pfohl, Agata Foryciarz, and N. Shah. 2021. An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction. Journal of biomedical informatics (2021), 103621.

[195] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. 2002. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in medicine 21 19 (2002), 2917–30.

[196] Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M. Jonker. 2011. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. User Modeling and User-Adapted Interaction 22 (2011), 357–397.

[197] William H. Press. 2009. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. Proceedings of the National Academy of Sciences 106 (2009), 22387 – 22392.

[198] Miao Qi, Owen Cahan, Morgan Foreman, Daniel Gruen, A. K. Das, and Kathleen P. Bennett. 2021. Quantifying representativeness in randomized clinical trials using machine learning fairness metrics. In medRxiv.

[199] M. Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Z. Wu. 2018. The Externalities of Exploration and How Data Diversity Helps Exploitation. ArXiv abs/1806.00543 (2018).

[200] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. New England Journal of Medicine 380, 14 (2019), 1347–1358.

[201] Alvin Rajkomar, Michaela Hardt, M. Howell, Greg Corrado, and M. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. Annals of Internal Medicine 169 (2018), 866–872.

[202] Meghna Ranganathan and Raj S Bhopal. 2006. Exclusion and Inclusion of Nonwhite Ethnic Minority Groups in 72 North American and European Cardiovascular Cohort Studies. PLoS Medicine 3 (2006).

[203] John Rawls. 1971. A Theory of Justice: Revised Edition.

[204] Shelby D Reed, Jui-Chen Yang, Timothy Rickert, F. Reed Johnson, Juan Marcos Gonzalez, Robert J. Mentz, Mitchell W. Krucoff, Sreekanth Vemulapalli, Philip B. Adamson, David J. Gebben, Liliana Rincon-Gonzalez, Anindita Saha, Daniel E Schaber, Kenneth M. Stein, Michelle E Tarver, and Dean Bruhn-Ding. 2021. Quantifying Benefit-Risk Preferences for Heart Failure Devices: A Stated-Preference Study. Circulation. Heart failure (2021), CIRCHEARTFAILURE121008797.

[205] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie Jane Calvert, Hutan Ashrafian, Andrew Beam, Gary S. Collins, Ara Darzi, Jonathan J. Deeks, Mohammed K. Elzarrad, Cyrus Espinoza, Andre Esteva, Livia Faes, Lavinia Ferrante di Ruffano, J Fletcher, Robert M. Golub, Hugh Harvey, Charlotte J Haug, Christopher Holmes, Adrian Jonas, Pearse A. Keane, Christopher J. Kelly, Aaron Y. Lee, Cecilia S. Lee, Elaine Manna, James Matcham, Melissa D. McCradden, David Moher, Joao Monteiro, Cynthia D. Mulrow, Luke Oakden-Rayner, Dina N. Paltoo, Maria Beatrice Panico, Gary Price, Samuel d. Rowley, Richard Savage, Rupa Sarkar, Sebastian J. Vollmer, and Christopher Yau. 2020. Guidelines for clinical trial protocols for interventions involving artificial

intelligence: the SPIRIT-AI Extension. The BMJ 370 (2020).

[206] D. Robertson, K. M. Lee, Boryana C. Lopez-Kolkovska, and S. Villar. 2020. Response-adaptive randomization in clinical trials: from myths to practical considerations. arXiv: Methodology (2020).

[207] David S Robertson, Babak Choodari-Oskooei, Munya Dimairo, Laura Flight, Philip Pallmann, and Thomas Jaki. 2021. Point estimation for adaptive trial designs. arXiv preprint arXiv:2105.08836 (2021).

[208] William F Rosenberger and John M Lachin. 2015. Randomization in clinical trials: theory and practice. John Wiley & Sons.

[209] William F Rosenberger, Diane Uschner, and Yanying Wang. 2019. Randomization: The forgotten component of the randomized clinical trial. Statistics in medicine 38, 1 (2019), 1–12.

[210] Peter M. Rothwell. 1995. Can overall results of clinical trials be applied to all patients? The Lancet 345 (1995), 1616–1619.

[211] Peter M. Rothwell, Ziyah Mehta, Sally C. Howard, Sergei A. Gutnikov, and Charles P. Warlow. 2005. From subgroups to individuals: general principles and the example of carotid endarterectomy. The Lancet 365 (2005), 256–265.

[212] Richard M. Royall, Robert H. Bartlett, Richard G. Cornell, David P. Byar, William D. Dupont, Robert J. Levine, Foster Lindley, R John Simes, and Marvin Zelen. 1991. Ethics and statistics in randomized clinical trials. Statistical science : a review journal of the Institute of Mathematical Statistics 6 1 (1991), 52–88.

[213] Rita Rubin. 2021. Pregnant People's Paradox-Excluded From Vaccine Trials Despite Having a Higher Risk of COVID-19 Complications. JAMA (2021).

[214] Janice Sabin, Frederick P. Rivara, and Anthony G Greenwald. 2008. Physician Implicit Attitudes and Stereotypes About Race and Quality of Medical Care. Medical Care 46 (2008), 678–685.

[215] Sara Saturni, Federico Bellini, Fulvio Braido, Pierluigi Paggiaro, Alessandro Sanduzzi, Nicola Scichilone, Pierachille Santus, Luca Morandi, and Alberto Papi. 2014. Randomized Controlled Trials and real life studies. Approaches and methodologies: a clinical point of view. Pulmonary pharmacology & therapeutics 27 2 (2014), 129–38.

[216] Scott Brian Saxman. 2015. Ethical considerations for outcome-adaptive trial designs: a clinical researcher's perspective. Bioethics 29, 2 (2015), 59–65.

[217] K. Schulman, J. Berlin, W. Harless, J. Kerner, S. Sistrunk, B. Gersh, R. Dubé, C. K. Taleghani, J. E. Burke, S. Williams, J. Eisenberg, and J. Escarce. 1999. The effect of race and sex on physicians' recommendations for cardiac catheterization. The New England journal of medicine 340 8 (1999), 618–26.

[218] David A. Schwartz. 2017. The Origins and Emergence of Zika Virus, the Newest TORCH Infection: What's Old Is New Again. Archives of pathology & laboratory medicine 141 1 (2017), 18–25.

[219] David A. Schwartz. 2018. Clinical Trials and Administration of Zika Virus Vaccine in Pregnant Women: Lessons (that Should Have Been) Learned from Excluding Immunization with the Ebola Vaccine during Pregnancy and Lactation. Vaccines 6 (2018).

[220] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency (2019).

[221] Anando Sen, Patrick B. Ryan, Andrew Goldstein, Shreya Chakrabarti, Shuang Wang, Eileen Koski, and Chunhua Weng. 2017. Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials. Annals of the New York Academy of Sciences 1387 (2017).

[222] Stephen J Senn. 2013. Seven myths of randomisation in clinical trials. Statistics in medicine 32 9 (2013), 1439–50.

[223] Sonia Mahesh Shah. 2003. Globalization of Clinical Research by the Pharmaceutical Industry. International Journal of Health Services 33 (2003), 29 – 36.

[224] Cong Shen, Zhiyang Wang, Sofia Villar, and Mihaela Van Der Schaar. 2020. Learning for dose allocation in adaptive clinical trials with safety constraints. In International Conference on Machine Learning. PMLR, 8730–8740.

[225] Janet E. Shepherd. 1994. Women and Health Research: Ethical and Legal Issues of Including Women in Clinical Studies. JAMA 272 (1994), 1467–1467.

[226] Richard Simon and Noah Robin Simon. 2011. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. Statistics & probability letters 81, 7 (2011), 767–772.

[227] Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. 2020. Racial bias in pulse oximetry measurement. New England Journal of Medicine 383, 25 (2020), 2477–2478.

[228] Vikas Soekhai, Chiara Whichello, Bennett Levitan, Jorien Veldwijk, Cathy Anne Pinto, Bas Donkers, Isabelle Huys, Eline van Overbeeke, Juhaeri Juhaeri, and Esther W. de Bekker-Grob. 2019. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug discovery today 24 7 (2019), 1324–1331.

[229] Harriette G. C. Van Spall, Andrew Toren, Alex Kiss, and R. A. Fowler. 2007. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 297 11 (2007), 1233–40.

[230] Francesco Sparano, Neil K. Aaronson, Francesco Cottone, Alfonso Piciocchi, Edoardo La Sala, Amélie Anota, Nina Deliu, Jacobien M Kieffer, and Fabio Efficace. 2019. Clinician-reported symptomatic adverse events in cancer trials:

are they concordant with patient-reported outcomes? *Journal of comparative effectiveness research* 8 5 (2019), 279–288.

[231] Tania Stafinski, Devidas Menon, Alexa A Nardelli, Jeffrey A Bakal, Justin A. Ezekowitz, Wayne Tymchak, Robert C. Welsh, Gabor T. Gyenes, and Paul W. Armstrong. 2015. Incorporating patient preferences into clinical trial design: results of the opinions of patients on treatment implications of new studies (OPTIONS) project. *American heart journal* 169 1 (2015), 122–31.e22.

[232] Nigel Stallard, Lisa Hampson, Norbert Benda, Werner Brannath, Thomas Burnett, Tim Friede, Peter K Kimani, Franz Koenig, Johannes Krisam, Pavel Mozgunov, et al. 2020. Efficient adaptive designs for clinical trials of interventions for COVID-19. *Statistics in Biopharmaceutical Research* 12, 4 (2020), 483–497.

[233] Kenneth E Stanley. 2007. Design of Randomized Controlled Trials. *Circulation* 115 (2007), 1164–1169.

[234] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. 2004. Randomized controlled trials. *American Journal of Roentgenology* 183, 6 (2004), 1539–1544.

[235] T Howard Stone. 2003. The invisible vulnerable: The economically and educationally disadvantaged subjects of clinical research. *Journal of Law, Medicine & Ethics* 31, 1 (2003), 149–153.

[236] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. 2009. Subgroup Analysis via Recursive Partitioning. *Econometrics: Single Equation Models eJournal* (2009).

[237] Oleksandr Sverdlov, Yevgen Ryeznik, and Weng Kee Wong. 2021. Opportunity for efficiency in clinical development: An overview of adaptive clinical trial designs and innovative machine learning tools, with examples from the cardiovascular field. *Contemporary clinical trials* (2021), 106397.

[238] Melanie M Taylor, Loulou Kobeissi, Caron Kim, Avni Amin, Anna Ekeus Thorson, Nita B Bellare, Vanessa Brizuela, Mercedes Bonet, Edna Oliveira Kara, Soe Soe Thwin, Hamsadvani Kuganantham, Moazzam Ali, Olufemi T. Oladapo, and Nathalie Jeanne Nicole Broutet. 2020. Inclusion of pregnant women in COVID-19 treatment trials: a review and global call to action. *The Lancet. Global Health* 9 (2020), e366 – e371.

[239] Fabio Alburquerque Thiers, Anthony J. Sinskey, and Ernst R. Berndt. 2008. Trends in the globalization of clinical trials. *Nature Reviews Drug Discovery* 7 (2008), 13–14.

[240] William R. Thompson. 1933. ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES. *Biometrika* 25 (1933), 285–294.

[241] Stefan Timmermans and Tara McKay. 2009. Clinical trials as treatment option: bioethics and health care disparities in substance dependency. *Social science & medicine* 69 12 (2009), 1784–90.

[242] Vasile Tiple. 2020. Recommendations on the European Commission's WHITE PAPER on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65 final (the 'AI White Paper'). *ISN: Other Legal Perspectives in Information Systems (Topic)* (2020).

[243] David J. Torgerson and Bonnie Sibbald. 1998. Understanding controlled trials: What is a patient preference trial? *BMJ* 316 (1998), 360.

[244] RJ Turner, A Ly, and PD Grünwald. 2021. Two-sample tests that are safe under optional stopping. (2021).

[245] Joseph M. Unger, Elise D. Cook, Eric W Tai, and Archie Bleyer. 2016. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting* 35 (2016), 185–98.

[246] Joseph M. Unger, Riha Vaidya, Kathy S Albain, Michael L. Leblanc, Lori M. Minasian, Carolyn Gotay, N. Lynn Henry, Michael J Fisch, Shing M. Lee, Charles D Blanke, and Dawn L. Hershman. 2022. Sex Differences in Risk of Severe Adverse Events in Patients Receiving Immunotherapy, Targeted Therapy, or Chemotherapy in Cancer Clinical Trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* (2022), JCO2102377.

[247] Berk Ustun, Yang Liu, and David C. Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In *ICML*.

[248] Rieke van der Graaf, Kit C. B. Roes, and Johannes J. M. van Delden. 2012. Adaptive trials in clinical research: scientific and ethical issues to consider. *JAMA* 307 22 (2012), 2379–80.

[249] Catharina M van Leersum, Albine Moser, Ben van Steenkiste, Marion Reinartz, Esther Stoffers, Judith R Wolf, and Trudy van der Weijden. 2020. What matters to me – a web-based preference elicitation tool for clients in long-term care: a user-centred design. *BMC Medical Informatics and Decision Making* 20 (2020).

[250] Eline van Overbeeke, Rosanne Janssens, Chiara Whichello, Karin Schölin Bywall, Jenny Sharpe, Nikoletta Nikolenko, Berkeley S. Phillips, Paolo Guiddi, Gabriella Pravettoni, Laura Vergani, Giulia Marton, Irina Cleemput, Steven R A Simoens, Jürgen Kübler, Juhaeri Juhaeri, Bennett Levitan, Esther W. de Bekker-Grob, Jorien Veldwijk, and Isabelle Huys. 2019. Design, Conduct, and Use of Patient Preference Studies in the Medical Product Life Cycle: A Multi-Method Study. *Frontiers in Pharmacology* 10 (2019).

[251] Michelle van Ryn and Julia Burke. 2000. The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social science & medicine* 50 6 (2000), 813–28.

[252] Yogatheesan Varatharajah, Brent M. Berry, Oluwasanmi Koyejo, and Ravishankar Krishnan Iyer. 2018. A Contextual-bandit-based Approach for Informed Decision-making in Clinical Trials. *ArXiv* abs/1809.00258 (2018).

[253] Jason L. Vassy, Yuk-Lam Anne Ho, Jacqueline Honerlaw, Kelly Cho, J. Michael Gaziano, Peter W. F. Wilson, and David R. Gagnon. 2018. Yield and bias in defining a cohort study baseline from electronic health record data. *Journal of biomedical informatics* 78 (2018), 54–59.

[254] Janette Vazquez, Samir E. Abdelrahman, Loretta M. Byrne, Michael Russell, Paul A. Harris, and Julio C. Facelli. 2020. Using supervised machine learning classifiers to estimate likelihood of participating in clinical trials of a de-identified version of ResearchMatch. *Journal of Clinical and Translational Science* 5 (2020).

[255] José Villar, Shabina Ariff, Robert B Gunier, Ramachandran Thiruvengadam, Stephen Rauch, Alexey Kholin, Paola Roggero, Federico Prefumo, Marynéa Silva do Vale, Jorge Arturo Cardona-Pérez, Nerea Maiz, Irene Cetin, Valeria M Savasi, Philippe Deruelle, Sarah Rae Easter, Joanna Sichitiu, Constanza P Soto Conti, Ernawati Ernawati, Mohak Mhatre, Jagjit S. Teji, Becky Liu, Carola Capelli, Manuela Oberto, Laura Salazar, Michael G. Gravett, Paolo Ivo Cavoretto, Vincent Bizor Nachinab, Hadiza S. Galadanci, Daniel Oros, Adejumoke Idowu Ayede, Loïc Sentilhes, Babagana Bako, Mónica Savorani, Hellas Cena, Perla K. García-May, Saturday Job Etuk, Roberto Casale, Sherief M. Abd-Elsalam, Satoru Ikenoue, Muhammad Baffah Aminu, Carmen Vecciarelli, Eduardo Alfredo Duro, Mustapha Ado Usman, Yetunde Olufisayo John-Akinola, Ricardo Nieto, Enrico Ferrazi, Zulfiqar A. Bhutta, Ana Ines Langer, Stephen H Kennedy, and Aris T. Papageorghiou. 2021. Maternal and Neonatal Morbidity and Mortality Among Pregnant Women With and Without COVID-19 Infection. *JAMA Pediatrics* 175 (2021), 1 – 10.

[256] S. Villar, J. Bowden, and J. Wason. 2015. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical science : a review journal of the Institute of Mathematical Statistics* 30 2 (2015), 199–215.

[257] Sofía S Villar, Jack Bowden, and James Wason. 2018. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical statistics* 17, 2 (2018), 182–197.

[258] Sofía S Villar, David S Robertson, and William F Rosenberger. 2021. The temptation of overgeneralizing response-adaptive randomization. *Clinical Infectious Diseases* 73, 3 (2021), e842–e842.

[259] Sofía S. Villar, James M. S. Wason, and Jack Bowden. 2015. Patient drift and response-adaptive randomisation: impact and solutions. *Trials* 16 (2015), P232 – P232.

[260] Sofía S. Villar, James M. S. Wason, and Jack Bowden. 2015. Response-adaptive randomization for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics* 71 4 (2015), 969–78.

[261] Darshali A Vyas, L. Eisenstein, and David S Jones. 2020. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England journal of medicine* 383 9 (2020), 874–882.

[262] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Amer. Statist. Assoc.* 113 (2018), 1228 – 1242.

[263] Karin A Wasmann, Pieta C Wijsman, Susan van Dieren, Willem A Bemelman, and Christianne Buskens. 2019. Partially randomised patient preference trials as an alternative design to randomised controlled trials: systematic review and meta-analyses. *BMJ Open* 9 (2019).

[264] James A. Watson and Chris C. Holmes. 2020. Machine learning analysis plans for randomised controlled trials: detecting treatment effect heterogeneity with strict control of type I error. *Trials* 21 (2020).

[265] Griffin M. Weber, William G. Adams, Elmer V. Bernstam, Jonathan P. Bickel, Kathe P. Fox, Keith A. Marsolo, Vijay A. Raghavan, Alexander Turchin, Xiaobo Zhou, Shawn N. Murphy, and Kenneth D. Mandl. 2017. Biases introduced by filtering electronic health records for patients with "complete data". *Journal of the American Medical Informatics Association* 24 (2017), 1134–1141.

[266] Noel S Weiss, Thomas D. Koepsell, and Bruce M. Psaty. 2008. Generalizability of the results of randomized trials. *Archives of internal medicine* 168 2 (2008), 133–5.

[267] E. Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F. Freitag, James Benoit, Michael C. Hughes, Faisal Khan, Paul Slater, Khader Shameer, Matthew Roe, Emmette R. Hutchison, Scott H. Kollins, Uli C. Broedl, Zhaoling Meng, Jennifer L. Wong, Lesley Curtis, Erich S Huang, and Marzyeh Ghassemi. 2021. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials* 22 (2021).

[268] David Wendler, Raynard S. Kington, Jennifer H. Madans, Gretchen Van Wye, Heidi Christ-schmidt, Laura A. Pratt, Otis W Brawley, Cary P. Gross, and Ezekiel J. Emanuel. 2006. Are Racial and Ethnic Minorities Less Willing to Participate in Health Research? *PLoS Medicine* 3 (2006).

[269] Graham M Wheeler, Adrian P Mander, Alun Bedding, Kristian Brock, Victoria Cornelius, Andrew P Grieve, Thomas Jaki, Sharon B Love, Lang'o Odondi, Christopher J Weir, et al. 2019. How to design a dose-finding study using the continual reassessment method. *BMC medical research methodology* 19, 1 (2019), 1–15.

[270] J. Wiens, S. Saria, Mark P. Sendak, M. Ghassemi, V. Liu, Finale Doshi-Velez, K. Jung, K. Heller, David C. Kale, Mohammed Saeed, P. Ossorio, Sonoo Thadaney-Israni, and A. Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* (2019), 1–4.

[271] David R Williams and R. Wyatt. 2015. Racial Bias in Health Care and Health: Challenges and Opportunities. *JAMA* 314 6 (2015), 555–6.

[272] J. J. Williams, Jacob Nogas, N. Deliu, Hammad Shaikh, S. Villar, Audrey Durand, and Anna N. Rafferty. 2021. Challenges in Statistical Analysis of Data Collected by a Bandit Algorithm: An Empirical Exploration in Applications to Adaptively Randomized Experiments. *ArXiv* abs/2103.12198 (2021).

[273] S. Williamson, P. Jacko, S. Villar, and T. Jaki. 2017. A Bayesian adaptive design for clinical trials in rare diseases. *Computational statistics & data analysis* 113 (2017), 136 – 153.

[274] K K Woloshin, Mack T. Ruffin, and Daniel W. Gorenflo. 1994. Patients' interpretation of qualitative probability statements. *Archives of family medicine* 3 11 (1994), 961–6.

[275] Marcus Y. Woo. 2019. An AI boost for clinical trials. *Nature* 573 (2019), S100–S102.

[276] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. 2021. A unified framework for bandit multiple testing. *Advances in Neural Information Processing Systems* 34 (2021).

[277] Fanny Yang, Aaditya Ramdas, Kevin Jamieson, and Martin J Wainwright. 2017. A framework for Multi-A (rmed)/B (andit) Testing with Online FDR Control. *arXiv preprint arXiv:1706.05378* (2017).

[278] Wei Yang, Nicole M Warrington, Sara J Taylor, Paula Whitmire, Eduardo Carrasco, Kyle W Singleton, Ningying Wu, Justin D Lathia, Michael E Berens, Albert H Kim, et al. 2019. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. *Science translational medicine* 11, 473 (2019).

[279] Q Yao and LJ Wei. 1996. Play the winner for phase II/III clinical trials. *Statistics in medicine* 15, 22 (1996), 2413–2423.

[280] M. B. Zafar, Isabel Valera, M. Gomez-Rodriguez, and K. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *Proceedings of the 26th International Conference on World Wide Web* (2017).

[281] M. B. Zafar, Isabel Valera, M. Gomez-Rodriguez, and K. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

[282] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[283] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*.

[284] R. Zemel, Ledell Yu Wu, Kevin Swersky, T. Pitassi, and C. Dwork. 2013. Learning Fair Representations. In *ICML*.

[285] B. Zhang, B. Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018).

[286] Kevin B. Zhang and Dina Demner-Fushman. 2017. Automated classification of eligibility criteria in clinical trials to facilitate patient-trial matching for specific patient populations. *Journal of the American Medical Informatics Association* 24 (2017), 781–787.

[287] Xingyao Zhang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2020. DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. *ArXiv* abs/2001.08179 (2020).

[288] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, et al. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet* 395, 10229 (2020), 1054–1062.

[289] Yurdaguel Zopf, Christina Rabe, Antje Neubert, Karl-Günter Gaßmann, Wolfgang Rascher, Eckhart G. Hahn, K. Prof.Dr.Chem. Brune, and Harald Dormann. 2008. Women encounter ADRs more often than do men. *European Journal of Clinical Pharmacology* 64 (2008), 999–1004.

[290] Irving Zucker and Brian J. Prendergast. 2020. Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences* 11 (2020).

A SEVEN PRINCIPLES FOR ETHICAL CLINICAL TRIALS

In discussions of ethical guidelines for clinical trials, three documents have been cited as universally influential [2, 76, 179]: the *Nuremberg Code* (1947) [55], the *Declaration of Helsinki* (1964) [11], and the *Belmont Report* (1979) [33], which were all developed following egregious instances of patient abuse. The Nuremberg Code is a response to the atrocities committed by Nazi doctors during World War II. It establishes the need for informed patient consent and a favorable risk-benefit ratio [55]. The Declaration of Helsinki was developed to supplement the Nuremberg Code and focuses upon the conduct between physicians and patients during research, particularly with respect to favorable risk-benefit ratio and independent review [76, 179].

Other notable documents [2] include the CIOMS (Council for International Organizations of Medical Sciences) *International Ethical Guidelines for Biomedical Research Involving Human Subjects* (2002) [84] and the *U.S. Common Rule* (1981) [143]. As these guidelines were created in reaction to particular circumstances, they have been criticized as lacking generality and even being in conflict with one another [52, 76]. Emanuel et al. [76] have sought to integrate the principles discussed by these guidelines into a unified framework and argue that these principles apply universally and are consistent with philosophies of "how reasonable people would want to be treated". Below, we detail the seven requirements developed by Emanuel et al. [76].

(1) **Value**: A research study must provide value by contributing useful knowledge to society and improving health and well-being. This requires that the research also be reliable, generalizable, and widely shared.

(2) **Scientific validity**: Research must be conducted in a "methodologically rigorous manner" [76], using widely accepted methods, principles, and practices, and be practically feasible. Trials that compare different treatments must abide by the principle of *clinical equipoise*, wherein researchers must possess "genuine uncertainty" regarding the benefits of new therapies over existing standard practices [85, 158].

(3) **Fair subject selection**: Participants should be selected for trials on the basis of the scientific goals of the study, rather than unrelated traits, such as vulnerability or privilege. Equally, certain groups or individuals may not be excluded without valid scientific reason or excessive risk. Those who may incur risk or burden of the research should be able to benefit, and those who may benefit should take on some of the risks or burdens.

(4) **Favorable risk-benefit ratio**: In clinical research, possible risks should be minimized, potential benefits should be maximized, and the potential benefits to study participants and to society should either be proportional to or outweigh the possible risks.

(5) **Independent review**: To minimize biases or competing interests, proposals for clinical research should be evaluated by independent bodies who are not affiliated with or influenced by those conducting the study.

(6) **Informed consent**: Individuals should be able to make an independent and well-informed decision, consistent with their own

values and preferences, regarding whether they want to participate in clinical research. To achieve this, individuals should i) be "accurately informed of the purpose, methods, risks, benefits, and alternatives to the research", ii) understand the provided information and its relevance to their own circumstances, and iii) be able to make a "voluntary and uncoerced" decision regarding whether or not to participate [76].

(7) **Respect for enrolled subjects**: Individuals should be treated with respect throughout the entirety of the clinical research process, beginning from recruitment and extending until after participation. This includes, but is not limited to the following: i) respecting their privacy, ii) respecting the right for them to change their minds and withdraw from participation, iii) informing participants of any new information regarding the interventions or their own clinical conditions, iv) monitoring their well-being and ensuring proper treatment throughout the trial process, and v) informing them of the results and impacts of the clinical research.

We note that the discussion of ethics surrounding clinical trials extends beyond the debate over how to conduct an ethical trial. Scholars continue to contemplate the intrinsic morality of the practice of clinical trials as a form of experimentation that is simultaneously viewed by participants as an opportunity for treatment. This is often without the full understanding of the research subjects, who may view clinical trials as their best or only opportunity for care [77, 241].

B BACKGROUND ON FAIRNESS IN MACHINE LEARNING

Anti-discrimination laws in the United States have provided two well-known interpretations for unfairness: *disparate treatment* and *disparate impact* [18]. *Disparate treatment* refers to the different treatment, with intent, of similar people (with respect to non-sensitive attributes, but different sensitive attributes), while *disparate impact* occurs with policies or practices that appear neutral but have a disproportionately adverse impact on those with certain sensitive attributes [18]. These notions lead to fairly straightforward translations into statistical definitions of group fairness.

B.1 Definitions of fairness

Group Fairness mostly measures the parity of a statistical measure (usually depending on the model outcomes and true outcomes) across all subgroups with different protected attributes [50]. For example, in the case of binary decisions, one can ask for equal rates of positive outcomes unconditionally of the true outcome (*demographic/statistical parity* or *equal allocation*). Conditioning on the true outcome yields a variety of definitions for so-called *classification parity*, for example equal rates of errors such as false positive and negative rates (*equal opportunity*, *equalized odds*, *disparate mistreatment*) [103, 280]. Conditioning on the predicted score instead, one can analogously define *calibration parity* by asking individuals from different groups with the same predicted score to have the same probability of actually achieving a positive outcome [58, 142, 280]. Generally, all combinations of false positive, negative, discovery, omission rates as well as positive or negative predictive values can be considered as meaningful fairness

definitions [282]. Such statistical group fairness notions are popular because they are straightforward to interpret and often to achieve during model training without making additional assumptions about the data generating process. However, a drawback is that they do not give meaningful guarantees to individuals, structured subgroups, or intersections of protected groups, but only to "average" members [50]. In addition, most subsets of this collection of definitions cannot be satisfied simultaneously [49, 142], leaving open the question of which criterion to choose. Instead, interventional notions of fairness take into consideration the causal structure, i.e., how the protected attribute may have influenced other features, and asks for equal decisions "had the protected attribute been fixed externally keeping everything else equal" [137]. Due to the assumption that the causal model [192] is known and ontological difficulties with attributing causal powers to variables such as "race" or "gender", these definitions are near impossible to operationalize [117].

Rajkomar et al. [201] propose three principles inspired by distributive justice [203] specifically considering health equity that resemble statistical group fairness notions: i) *equal patient outcomes* (when all subgroups receive equal benefit from the model), ii) *equal performance* (when a model is equally accurate for all subgroups), and iii) *equal allocation* (when resources are proportionately allocated to patients in all subgroups). *Equal patient outcomes* is difficult to analyze as downstream effects of decisions may be unpredictable. While *equal performance* and *equal allocation* can be formalized as statistical group fairness criteria, they do not necessarily translate to *equal outcomes*. While *equal allocation* (statistical parity) is considered a crude criterion that is entirely blind to true outcomes, it can be relevant in healthcare settings. For example, historically, African American women with chest pain were sent for cardiac catheterization treatment at a lower rate in comparison with white men [217], such that *equal accuracy* would still propagate inequality, as these patients would be under-identified.

Individual fairness attempts to capture injustice experienced by a single individual compared to other (similar) individuals with different protected attributes. Dwork et al. [74] accordingly suggest that an algorithm is fair if similar individuals, according to a task-specific metric on the inputs, receive similar (distributions over) outcomes, again according to a fixed metric on outcomes. The main drawback is that the choice of similarity metrics is non-trivial in that they ultimately have to capture in which regards people should be considered similar/equal [50]. Despite attempts to learn such a metric from various types of online feedback [22, 93, 118, 177], individual fairness remains difficult to implement in practice. *Counterfactual fairness* may be viewed as a form of individual fairness that defines similarity with respect to a causal model of all measured features and protected attributes [144]. The underlying idea is that an individual should receive the same outcome in a counterfactual world in which they had a different protected attribute all else being equal. Again, due to the strong assumption of knowing the causal model, counterfactual fairness suffers from the same drawbacks as interventional fairness.

Another common definition. that fits to neither of the two categories is *fairness through unawareness* (or *anti-classification*),

wherein an algorithm is fair if it does not consider sensitive attributes during its decision making process [58, 171]. However, this practice has shown to be ineffective and potentially harmful [17, 58].

B.2 Approaches to achieve fairness

Techniques to achieve fairness in ML models are commonly categorized by when the intervention occurs in the model building pipeline: pre-processing, at training time, and post-processing [17, 171].

Pre-processing. In the pre-processing approach, practitioners seek to transform the feature space into a representation that is independent of the sensitive attribute. This approach is agnostic to the downstream tasks that accept the representation as input. Correcting data biases is a difficult task, as it requires an understanding of how the measurement process is biased or intuition about how the data would appear in an "unbiased" setting [50]. Recent work has proposed that model prediction error can be decomposed in terms of bias, variance, and noise, and that these values can be used to inform additional data collection [42]. Practically, learning fair representations is predominantly based on two potentially competing objectives: maintain information in the inputs that is relevant for accurate decision-making while removing all information about the protected attribute of a given input [80, 166, 284, 285].

At Training Time. At training time, fairness can be included as a constraint on the loss minimization. Achieving independence from the sensitive attributes at training time is beneficial as the classifier can be optimized with the specific fairness criterion in mind; a drawback is that we require access to the model training pipeline, and the final approach is usually model- or problem-specific [17]. Typically, this approach involves either changes of the objective function or imposing a constraint based on the previously discussed fairness definitions [17, 171, 281].

Post-Processing. In post-processing, a practitioners adjusts (the outputs of) a trained model so that it achieves a desired fairness criterion [17, 171]. Post-processing is most useful when one does not have access to the training data or learning algorithm. Hardt et al. [103] introduce a method to achieve equalized odds or equality of opportunity by (stochastically) changing certain model outputs. Similarly, Kim et al. [139] propose a framework to audit and post-process trained models to ensure accurate predictions across specified population subgroups.