Philosophy of Medicine
Analysis

Randomized Controlled Trials in Medical AI
A Methodological Critique[1]

Konstantin Genin
Research Group: "Epistemology and Ethics of Machine Learning"; Cluster of Excellence: Machine Learning: New Perspectives for Science; University of Tübingen, Germany
Email: konstantin.genin@gmail.com

Thomas Grote
Ethics and Philosophy Lab; Cluster of Excellence: Machine Learning: New Perspectives for Science; University of Tübingen, Germany
International Center for Ethics in the Sciences and Humanities (IZEW); University of Tübingen, Germany
Email: thomas.grote@uni-tuebingen.de

Various publications claim that medical AI systems perform as well, or better, than clinical experts. However, there have been very few controlled trials and the quality of existing studies has been called into question. There is growing concern that existing studies overestimate the clinical benefits of AI systems. This has led to calls for more, and higher-quality, randomized controlled trials of medical AI systems. While this a welcome development, AI RCTs raise novel methodological challenges that have seen little discussion. We discuss some of the challenges arising in the context of AI RCTs and make some suggestions for how to meet them.

1. Introduction

Recent years have seen increased interest in the application of artificial intelligence (AI) for clinical decision-making. Various high-profile publications claim that medical AI systems perform as well, or better, than clinical experts—especially when diagnosing disease on the basis of medical images (see Topol 2019 for a review). Prominent examples include applications in dermatology (Liu, Jain et al. 2020), ophthalmology (Gulshan et al. 2016) and oncology (Esteva et al. 2017). Despite these developments, there are growing concerns that these studies overestimate the clinical benefits of AI systems in realistic settings. Most existing studies are retrospective and performed outside of clinical environments (Topol 2020). Moreover, outcomes used to evaluate AI performance tend to be only surrogates for meaningful clinical endpoints (Oren, Gersh and Bhatt 2020). Finally, only a handful of

[1] Joint first authorship.

ISSN 2692-3963 | Date Received 07-10-2020| Date Accepted 13-04-2021| Date First Published 28-05-2021
Philosophy of Medicine | DOI 10.5195/POM.2021.27 | Volume 2 | pp. 1-15

randomized clinical trials (RCTs) have been performed. In one of these few RCTs, Haotian Lin et. al (2019) fail to replicate the apparent AI advantage reported by Erping Long et al. (2017); they find that AI systems are *less* accurate than senior consultants. In their meta-analysis, Xiaoxuan Liu et al. (2019) found that many AI studies are not transparent about their design and methods, creating barriers to the interpretation and replication of their results. Lastly, most AI studies are framed antagonistically: algorithms are made to compete with clinicians in the hopes of demonstrating their superiority. In all likelihood, medical AI will ultimately supplement, and not replace, human judgement. The envisioned role of AI systems is typically that of a second reader, consulted by a clinician who is uncertain about her initial diagnosis (McKinney et al. 2020). That renders antagonistic studies uninformative about the clinical promise of medical AI (Topol 2020).

Critiques of research in clinical AI usually culminate in a call for more RCTs. The emerging standard is for RCTs with clearly reported designs and methods, in realistic clinical environments, with meaningful clinical endpoints and in which clinicians are assisted, but not replaced, by AI systems. In support of this emerging standard, existing reporting guidelines for clinical trials have been extended to provide guidance for clinical trials involving AI systems (Cruz Rivera et al. 2020; Liu, Cruz Rivera et al. 2020; Mongan, Moy and Kahn 2020). These extensions focus on improving transparency in reporting trial design and methodology, with a view to easy interpretation, review, and replication. Inspired by multi-phase clinical drug trials, Yoonyoung Park et al. (2020) sketch a methodology of staged trials for clinical AI. These are welcome developments. However, these extensions do not aim to be prescriptive about RCT methodology and refrain from discussions of the unique methodological issues that arise for AI RCTs. In this article, we attempt such a discussion.

RCTs are relatively well studied in the philosophy of medicine. In broad terms, the debate has centred around the privileged status of RCTs within the evidential hierarchy of medical research. RCTs are widely accepted as the methodological "gold standard"—they are taken to be the only reliable methods for establishing causal claims about the efficacy of a given intervention. Philosophers have contested this status, by questioning the rationale for randomization (Urbach 1985, 1993; Worrall 2002, 2007); by highlighting evidential constraints regarding the transferability of average treatment effects (see Cartwright 2007; Deaton and Cartwright 2018; Worrall 2010); or by arguing that in order to establish causal claims, RCTs need to be supplemented by mechanistic evidence (Russo and Williamson 2007). For the most part, we take for granted the methodological superiority of the RCT: randomization and control are necessary for arriving at a statistically unbiased estimate of the average treatment effect. However, randomization and control are not sufficient for this purpose. In this article, we address unique threats to validity arising in the context of AI RCTs. Many of these threats to validity arise post-randomization. Although Angus Deaton and Nancy Cartwright (2018) provide a useful catalogue of potential post-randomization biases, these problems have been relatively neglected by philosophers of science. A particular focus of this article is how post-randomization threats to validity arise in the context of AI RCTs.

The remainder of this article is structured as follows: Section 2 reviews the existing critiques of research on medical AI; section 3 reviews the few medical AI RCTs that have already been performed; and, finally, section 4 discusses some of the methodological challenges arising in the context of AI RCTs, with some suggestions for how to meet them.

2. The Need for Medical AI RCTs

Although computers have been assisting in clinical decisions since the 1970s (see Schaffner 1985; Berner and La Lande 2016; Varghese et al. 2018), advances in deep-learning-based computer vision have set off a new wave of research in AI-assisted diagnosis, prognosis, and treatment. Since Varun Gulshan et al. (2016) developed an algorithm to detect diabetic retinopathy from fundus images, many studies have announced AI systems diagnosing diseases at "expert level" (see Topol 2019). This, in turn, has fuelled the interest in using this latest generation of AI systems in clinical settings.

Most of these studies share a common design (Grote and Berens, forthcoming). First, an AI system (typically a deep neural network) is trained to classify different disease entities on the basis of medical images annotated by medical professionals. To validate the AI system, its accuracy is measured in a benchmark dataset, which has been labelled according to an external standard. If the algorithm performs well on the benchmark task, its performance is compared to clinical experts asked to classify the same images. If it performs similarly or better than the clinicians, it is considered validated. Although the performance of these systems is impressive, recent meta-analyses by Liu, Faes et al. (2019) and Myura Nagendran et al. (2020) raise concerns that these retrospective in-silico studies overestimate the clinical benefits of AI systems in realistic settings. We can summarize their main critiques as follows:

i. *Unfair comparison:* The validation task does not reflect the real diagnostic abilities of an expert clinician, as unlike in real clinical settings, she has no access to other diagnostic modalities (for example, patients' testimonies, patients' health records, medical devices).
ii. *Irrelevant comparison:* Many studies compare surrogate endpoints (diagnostic accuracy) whose bearing on ultimate clinical outcome is unclear. Identifying more diminutive adenomas may prevent cancer, or it may result in unnecessary surgery or chemotherapy. We discuss this in more detail in the fourth section. Moreover, if the algorithms are supposed to assist clinicians, employing an antagonistic study design in which the two compete is not really informative.
iii. *Unreliable comparison:* The studies compare the performance of AI with only small groups of clinical experts (with a median of only four).
iv. *Unrealistic comparison:* The studies are retrospective and do not take place in realistic clinical settings.
v. *Few RCTs:* Most of the studies are not randomized controlled trials and virtually none are double blind.
vi. *Bad reporting:* Most of the studies do not adhere to reporting standards. For instance, the training data, procedure, or algorithmic details are not made transparent.
vii. *Exaggerated claims:* Many claims about AI effectiveness in the relevant studies are not backed by their own statistical analyses.

These critiques raise concerns about the reproducibility and external validity of existing studies. If apparently validated AI systems turn out to be unreliable in clinical settings, the consequences are potentially disastrous.

The standard emerging from these critiques is for randomized and controlled trials, with clearly reported designs and methods, in realistic clinical environments, with meaningful clinical endpoints and in which clinicians are assisted, but not replaced, by AI systems. In support of this emerging standard, existing reporting guidelines for clinical trials (SPIRIT and CONSORT) have been extended to provide guidance for clinical trials involving AI

systems (Liu, Cruz Rivera et al. 2020; Mongan, Moy and Kahn 2020).[2] The focus of these guidelines is to improve transparency in reporting trial design and methodology, with the aim of facilitating the interpretation, review, and replication of the studies. To give some examples, researchers are encouraged to specify the model architecture of the algorithm and its training process. In addition, they are encouraged to upload their source code to a public repository. Furthermore, the researchers are required to be explicit about the AI-intervention, the clinical setting, and the modality of human-AI interaction. They are also directed to detail cases in which the AI failed or misled clinicians. There is little doubt that the authors of the extensions address many of the critiques of the meta-analyses. However, these extensions explicitly do not aim to be prescriptive about methodology and thus abstain from discussions of the unique methodological issues that arise for AI RCTs. In section 3, we attempt such a discussion. In preparation, we review the few RCTs that have been performed in this area.

3. **Randomized Controlled Trials for Clinical AI: The State of the Art**

In a recent commentary on the SPIRIT and CONSORT extensions, Eric J. Topol (2020) reports only seven completed RCTs involving AI-assisted treatment in prospective studies in real clinical environments. All but two studies (Wijnberge et al. 2020 and Lin et al. 2019) concern AI-assisted endoscopy. All but one (Wijnberge et al. 2020) were performed in four hospitals in China. Most studies had large numbers of patients, but small numbers of clinicians. One study enrolled four clinicians; three studies enrolled six clinicians, and one enrolled eight. In the case of Marije Wijnberge et al. (2020), we were not able to determine how many clinicians were involved. It was often unclear whether each clinician treated patients from only a single arm of the study, or whether a single clinician might treat patients from both arms. It was often unclear how the clinicians were assigned to experimental or control groups, or whether the groups were roughly equal in experience or whether their caseload (outside of the study) was roughly similar. Since the number of clinicians tends to be very small, reviewers should be rigorously aware of small sample size effects. All could improve their reporting in this respect.

Every study, except Lin et al. (2019), had a collaborative design: clinicians in the intervention group were assisted, not replaced, by AI. Although they chose an antagonistic design, the Lin et al. study is interesting because it fails to replicate an AI advantage found by Long et al. (2017) in a retrospective, antagonistic study. In fact, Lin et al. found that the AI system was even *less* accurate than senior consultants in diagnosing childhood cataracts.

The six collaborative studies can be divided into three that perform *technical* and three that perform *cognitive* interventions.[3] In technical studies, the AI system performs a role that would otherwise be performed by a piece of equipment. Dexin Gong et al. (2020), Lianlian Wu et al. (2020) and Wijnberge et al. (2020) perform technical interventions. Both Gong and Jing-Ran Su et al. (2020) develop systems to standardize colonoscopy quality by making the clinician slow down and avoid blind spots. Wijnberge et al. develop an early warning system for hypotension during surgery. These are technical improvements that otherwise might have been achieved by a sophisticated stopwatch or blood pressure monitor. In cognitive interventions, the AI system performs a role normally performed by an expert clinician. The AI system provides clinicians with diagnostic decisions/predictions, which then need to be weighed by the clinicians. Pu Wang et al. (2019, 2020) and Su et al.

[2] SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials (Artificial Intelligence); CONSORT: Consolidated Standards of Reporting Trials (Artificial Intelligence).
[3] This distinction is due to Gong et al. (2020).

(2020) perform such cognitive interventions. Their algorithms identify polyps and adenomas on the basis of endoscope images, a task normally performed by expert clinicians. These studies are the focus of our discussion because they pose the full range of methodological challenges that arise for AI RCTs. For example, Wang et al. (2020) worry that the physicians receiving algorithmic assistance might develop a "competitive spirit" (344) with respect to the algorithmic system and thereby distort the effects of the intervention. Although it is possible that these kinds of effects arise for technical interventions, it is much less plausible that physicians would develop competitive attitudes towards a stopwatch, blood pressure monitor, or other relatively *un-opinionated* device.

Only Wang et al. (2020) implemented a double-blind design. In this study, they attempt to replicate the results of their earlier (2019) single-blind study. Prima facie, it is difficult to blind clinicians to the presence of AI assistance: how could the clinician fail to notice if she received algorithmic support? To address this, Wang et al. implement a blinding strategy involving the use of a *sham* AI. Because of its ambitious design, Wang et al. represents the cutting edge of AI RCTs. It also poses the most interesting methodological problems. For this reason, it is a particular focus of our discussion below.

4. **Revisiting the Methodology of Medical AI RCTs**

Although the methodological superiority of RCTs is widely accepted, philosophers of science have questioned whether RCTs deserve the esteem in which they are held by methodologists.[4] Indeed, the methodological advantages of RCTs are rarely explained clearly and are subject to many misconceptions (see Senn 2013; Fuller, 2019). We take for granted that the goal of clinical trials is to arrive at statistically unbiased estimates of the average effect of the treatment under study.[5] Researchers might systematically overestimate the benefit of treatment if patients assigned to treatment are more likely to be young, healthy, or rich than those assigned to control. Random assignment facilitates unbiased estimation of average treatment effects by rendering baseline prognostic factors (such as age, race, disease severity, and so on) statistically independent of assignment to treatment. Although the details differ, all prominent formal frameworks for reasoning about probabilistic causal inference agree that statistical independence of baseline prognostic factors from assignment to treatment is necessary for unbiased inference and that random assignment is—at least sometimes—sufficient for ensuring that the independence holds.[6] Moreover, the usual rationale for randomization transfers fairly straightforwardly to AI RCTs: randomization ensures that the population of patients receiving AI-assisted care are not more likely to have baseline factors more favourable (or unfavourable) for the outcome than patients receiving regular care.

Statistical independence between baseline prognostic factors and assignment is necessary for unbiased inference, but it is not sufficient. A variety of biases can arise post-randomization. If, for example, random assignment to placebo makes it more likely that the attending clinician will prescribe some concomitant therapy as "compensation", then the trial may underestimate the direct benefit of treatment. These are the kinds of problems that blinding is meant to solve. Deaton and Cartwright (2018) provide a useful catalogue of common sources of post-randomization bias. Particularly relevant in our context are the

[4] Prominent critics include Urbach (1985, 1993) and Worrall (2002, 2007).
[5] In what follows, we will simply say "unbiased" for statistically unbiased, the property enjoyed by estimators that are equal, in expectation, to the parameter of interest.
[6] See Hernàn (2004), Steel (2011) and Deaton and Cartwright (2018) for perspicuous explanations of the virtues of randomization, couched in the different languages of competing formal frameworks for causal inference.

John Henry and physician effects. The former refers to the tendency of the control group to develop a competitive attitude towards the experimental group and thereby invalidate their status as a control. "John Henry" refers to the folkloric railroad worker who, upon learning that his performance would be compared to the steam drill, worked so hard to outperform the machine that he died. Although the consequences would not be as dire, it is easy to imagine physicians similarly motivated by a spirit of competition with diagnostic algorithms. The physician effect refers to the idiosyncratic effect of the individual physician on patient outcome, beyond that of the treatment under study. Neither of these effects are straightforwardly mitigated even by successful blinding. In the best case, blinding will temper a John Henry effect into a Hawthorne effect: subjects will modify their behaviour as a result of an awareness of being observed and compared, but the modification will not be preponderant in one or the other group. Although successful blinding removes the threat the John Henry effect poses to internal validity, a threat to external validity remains: it is difficult to predict the outcome of the treatment once the subjects are no longer aware of being observed. Moreover, blinding alone does not mitigate the physician effect: if, for example, experienced physicians are more likely to treat patients in the experimental group, we would expect the effectiveness of the treatment to be overestimated. We turn now to the ways in which physician effects arise in AI RCT trials.

The results of an AI RCT would be misleading if clinicians receiving AI assistance tended to be more experienced, specialized, or less burdened than those in the control group. We would expect patients treated by more experienced physicians to have better outcomes, regardless of AI assistance. Moreover, experience may interact with the treatment in unexpected ways: more experienced clinicians may react very differently to AI assistance than their less experienced colleagues. For example, Philipp Tschandl et al. (2020) studied dermatologists interacting with an image-based AI for diagnosing skin cancer. They found that less experienced clinicians tended to accept AI-based support that contradicted their initial diagnosis even if they were very confident. More experienced clinicians, by contrast, tended to change their diagnoses to agree with the AI only when they were not confident. Although less experienced clinicians benefited significantly from AI assistance, experienced clinicians benefited only marginally. However, in cases when they were antecedently confident about their diagnosis, experienced clinicians performed worse with AI support—in the rare cases when they changed their diagnoses to agree with the AI, they tended to be led astray. Assuming the results of Tschandl et al. are representative, AI trials would not be probative about the usefulness of AI assistance if senior clinicians were over-represented in the intervention group: treatment decisions would be hardly changed. Conversely, if junior clinicians were over-represented in the intervention group, the trial would overestimate the benefits of AI, since adverse effects on experienced clinicians would rarely be observed.

To adequately control for these potential physician effects, it is important for any study to be clear about the distribution of experience and expertise among clinicians in the experimental and control groups, especially since the number of clinicians involved tends to be very small. Random assignment can mitigate the effect of physician effects, but a lot depends on the details of randomization. For example, researchers could randomize each physician to either always or never receive AI assistance. The trouble with this scheme is that, in the usual case when only a few physicians are participating in the trial, small sample-size effects may predominate: if only three senior and three junior clinicians are participating in the trial, it is not unlikely that all the senior physicians are assigned to always receive AI support. Moreover, in such a scheme, it is not possible to compare an individual physician's performance with and without algorithmic assistance—only the average difference between patient groups can be measured. A more promising scheme first

assigns patients to physicians and then randomly assigns each unique physician-patient pair to treatment or control. This scheme ensures independence between clinician experience and treatment assignment and also enables analysts to compare individual physicians to themselves with and without AI assistance. In this way, each physician serves as her own control. Although this is relatively standard practice in drug trials, it is unclear if it is followed in existing AI RCTs. Only Wang et al. (2019) include comparative statistics on clinician expertise, and this is dropped in Wang et al. (2020). Investigators should be encouraged to be clear on this matter.

Although physician effects can be mitigated by appropriate randomization, John Henry and related effects must be dealt with in some other way. Nagendran et al. (2020) and Yuichi Mori, Shin-ei Kudo and Masashi Misawa (2020) call for increasing use of double-blind designs in AI RCTs. Double-blind AI RCTs are indeed rare: so far, only Wang et al. (2020) have performed such a study. The usual methodological justification for blinding the clinician is to ensure that preconceived ideas of the investigator are not important to patient outcomes (Friedman, Furberg and DeMets 2010). Of course, blinding does not remove the influence of preconceived ideas, but it does ensure that their effects are not preponderant in any single group. For example, if clinicians are hostile to AI assistance, they may unconsciously sabotage it. If they are uncritical boosters of AI, they may put more effort in when they receive AI assistance than they do without it. To ensure that these effects are not concentrated in the experimental or control group, the clinicians could be blinded to the use of AI assistance. This is prima facie difficult: how could you not know if you were receiving AI assistance? One way this could be achieved is with a Turing-style design. For example, in AI-assisted colonoscopy, the AI displays alerts for adenoma structures that appear in the visual field. In a Turing-style design a human clinician sitting in a separate room could generate the alerts instead. In such a design, the operating endoscopist would not know whether a human or a machine were generating the alert. Of course, the results of such a trial would only bear on how AI assistance compares with the human assistance and not on whether it is better than no assistance at all. For this reason, it may be desirable to run a three-way trial in which a third group of patients is randomized to receive care without additional assistance.

Wang et al. (2020) motivate their double-blind design by appeal to a kind of John Henry effect:

> One major limitation of the existing non-blinded studies was the introduction of operational bias, because operating endoscopists using the CADe system might be more vigilant because of a competitive spirit or relax and rely on the CADe system. In both cases, the effectiveness of the CADe system might be overestimated or underestimated. (344)

Wang et al. should be commended for their attention to these potential biases. However, their attempts to account for it are, in our opinion, counterproductive. To mask the endoscopists, Wang et al. developed a "sham system" that "simulated alert boxes on polyp-like non-polyp structures (e.g., bubbles, faeces, undigested debris, and wrinkled mucosa) without tracking actual polyps during the colonoscopy" (2020, 345). Then, the output of both the CADe and the sham system was shown on a second monitor, which was visible only to an observing senior endoscopist and not the operating endoscopist. In both groups, "the observer was responsible for reporting the location of any visible alert box for the endoscopist with a laser pointer on the primary screen." As the authors note, the very fact of being observed might have improved inspection technique and "motivated the

competitive spirit" of the operating endoscopists—that may go some way towards equalizing novelty effects across the two groups. However, endoscopists receiving AI assistance were compared to endoscopists distracted with irrelevant and deliberately misleading laser alerts. The tendency of this design is to exaggerate the helpfulness of AI assistance by comparing it, not with the absence of AI assistance, but with the presence of algorithmic sabotage. This sort of design is convoluted and counterproductive and it also raises ethical concerns, as it imposes unnecessary risks on patients. If a Turing-style design is impractical, endoscopists receiving AI assistance could have been compared with endoscopists stimulated by the supervision of a senior colleague. Of course, that would mean that AI-assisted treatment would have to pass a rather severe test: it would have to be an improvement over two clinicians working in tandem. In yet another approach, the AI would assist with *every* patient, but randomize how "helpful" it is going to be—for example, by reducing or increasing the number of interventions. In this way, all clinicians could be made to anticipate AI assistance. Then, endoscopy sessions receiving significant assistance could be compared to those receiving minimal assistance—all clinicians would then be motivated by (perceived) algorithmic competition. The benefits of the Wang et al. design could be had without distracting and misleading practising clinicians with flashing lights.

So far, we have been concerned with potential threats to *internal* validity. But even studies that provide an excellent, unbiased estimate of the average treatment effect in the trial population may fail to generalize outside of the context of the trial. In what follows, we consider what features of AI RCTs threaten their *external* validity. In the context of a trial, clinicians are interacting with an untested AI system. They may regard it critically or be moved to greater concentration by a spirit of competition. This may give a temporary and artificial advantage to clinicians in the experimental group. In a short trial, researchers will oversample the period in which clinicians are still adjusting to the new system and before they are able to use it effectively. They will not yet understand the AI system's basic capabilities and limitations, or its medical point of view: how severely it grades disease, or how to interpret its probabilistic output (Cai et al. 2019). Though present during the study, these novelty effects would subside outside of it. Were the AI to be widely adopted, clinicians would be interacting with a "proven" system. They would have become acclimated to the AI. In the long run, they may learn to ignore it, or they may be coaxed into an over-reliance on its assistance (Park et al. 2020). A straightforward approach to dealing with novelty effects is to wait until they wash out: if AI trials run longer than a few weeks, clinicians will have time to acclimate themselves to the new system. Then, comparisons of the experiment and control outcomes could investigate how differences in outcomes evolve over time. This would give the most realistic picture of how the effect of AI assistance would evolve in a clinical setting.

It is important to note that widespread adoption of AI systems may pre-empt the development of certain kinds of expertise. If junior clinicians are over-reliant on AI systems, they may never develop the mastery of their more senior colleagues. Junior clinicians may be inadvertently trained to uncritically imitate AI systems, instead of critically collaborating with them. In this way, AI systems may eventually be used by clinicians who are less confident, experienced, or critical than those on whom they were originally tested. These considerations argue in favour of long-run clinical trials that investigate both how AI assistance interacts with clinician experience and how these effects evolve over time. Over-reliance on AI assistance might be mitigated by algorithmic explanations. Tschandl et al. (2020) argue that explanations provided by the AI system (by way of a heatmap, which highlights regions of interest) play an important pedagogical role as clinicians progress from novice to expert (on explainable AI in medicine, see Erasmus, Brunet and Fisher 2020;

Sullivan, forthcoming). Through the explanations, the clinicians learn to direct their attention to meaningful signs and symptoms. Of course, this assumes that the AI's judgements—as well as its "reasons" are themselves reliable. In any case, it is better that clinicians understand the reasoning of the AI system, so that they can adjudicate any disputes with standard clinical reasoning. So far, medical AI RCTs have not investigated to what extent explainability improves the diagnostic support of the AI system. We believe that it is important to close this research gap, to get better evidence on when and what sort of explanations are required to improve the interplay of AI systems and clinicians.

Additionally, researchers should take care that the introduction of AI assistance does not induce a survival bias. For example, Google Health deployed the Gulshan et al. (2016) algorithm for detecting diabetic retinopathy in retinal photographs in eleven clinics in Thailand (Beede et al. 2020). To ensure accuracy, the system accepted only high-quality images. Since many images were taken in poor lighting conditions, more than a fifth were rejected. Patients with rejected images were asked to come back another day. Poor internet connections also caused problems with uploading the images. This study was not an RCT, but if it had been, it is easy to imagine how these problems would induce survival bias: patients at well-equipped clinics with stable internet connections would be over-represented in the experimental group. In all likelihood, richer patients would therefore also be over-represented, even if assignments were randomized. This would probably overestimate the effectiveness of AI assistance. AI trials should be vigilant with regard to these possibilities. If possible, they should include the intention-to-treat analysis as well as a per-protocol analysis. Researchers should justify their decisions to perform one or the other kind of analysis. The failure of Google Health to successfully implement the AI system also highlights some constraints with respect to the transferability of AI systems to different environments. While the relevant systems may work reliably in a *state-of-the-art* academic hospital, they may be useless for less well-equipped hospitals.

The choice of which clinical endpoint to measure and compare has proven to be intricate. Ideally, what you want to establish in an RCT is that a given treatment had a meaningful effect on patient health (for example, whether there is an increase in the survival rate or recovery time). The problem with current AI systems is that they are neither a treatment in themselves, nor do they determine treatment on their own. What they do instead is to provide a secondary diagnostic opinion to the clinician. The role of the AI system is therefore causally upstream of treatment (see Lalumera and Fanti 2019 for similar concerns regarding medical imaging technologies). The decision of the AI system could be ignored by the clinician, or otherwise be irrelevant for the choice of treatment. This makes it tempting to choose a surrogate endpoint, such as diagnostic accuracy, as in Wang et al. (2020). What speaks in favour of diagnostic accuracy is that by getting the diagnosis correct, it spares patients an odyssey of further diagnostic tests—which in itself can be considered as a quality of life mprovement. However, relying on a surrogate endpoint may backfire. A particular worry is that the involvement of AI systems leads to overtreatment (Oren, Gersh and Bhatt 2020). While an AI system may spot tumours more accurately than even expert clinicians, these previously overlooked tumours may be clinically irrelevant. Wang et al. (2019, 2020) find that their system increases the detection of small and diminutive polyps, whose relevance for colorectal cancer prevention is debatable (Vleugels et al. 2017). However, once spotted, further interventions that are harmful to the patient are highly likely to follow, from biopsy to chemotherapy. Hence, merely focussing on surrogate endpoints is insufficient to establish the medical benefit of AI systems. Theoretically, the problem might be mitigated by using a more refined metric of diagnostic accuracy, which includes parsing

which disease will impact patients, and which will not.[7] However, making these sorts of prognoses is beyond what current image-based diagnostic AI systems are capable of.

Although RCTs are necessary for testing the mettle of AI systems, they are not sufficient in and of themselves. A fundamental question any AI trial should be able to answer is how, if at all, AI assistance changed decisions about diagnosis or treatment. In drug trials, the answer is relatively simple: patients in one group were assigned to a drug regimen, while the others were assigned to placebo. To answer this question in the case of AI systems may require building models to predict treatment or diagnostic decisions in both arms of the study and comparing them for salient differences. If AI assistance improves patient outcomes, researchers should ensure that these improvements are stable across time, patient characteristics, and across clinicians of different specializations or levels of experience. It is not enough to demonstrate an improvement in patient outcomes: effort should be made to identify the mechanism by which the improvement was made. The latter point is fairly commonplace in the health sciences: Federica Russo and Jon Williamson (2007) argue that probabilistic evidence for causal conclusions in the health sciences must always be buttressed with plausible biomedical mechanisms. However, this point takes on a somewhat different aspect for AI RCTs. In an AI RCT study, we are not looking for a biological pathway, but an institutional and procedural pathway: how did interacting with the algorithmic system change diagnostic and therapeutic practice? Knowledge of biological mechanisms and anatomical microstructures cannot answer this question. Rather, researchers would have to combine statistical and sociological/ethnographic methods to understand the effect AI intervention has on medical practice.

Finally, even if the involvement of AI systems improves the health outcomes of patients, it could still be detrimental to the clinician-patient relationship (cf. Bjerring and Busch 2020; Grote and Berens 2020). For instance, if AI assistance speeds up the diagnostic process, this could come at the expense of the care component, even if speeding up the diagnostic process theoretically frees up resources for care work. Moreover, AI assistance might interfere with the trust relationship between the patient and the clinician. If the clinician tends to (over-)rely on diagnostic support, the patient may suspect that it is not the clinician, but the AI system that is in command. The bottom line is that while AI systems may contribute to the instrumental aim of medicine (curing disease), these algorithms may nevertheless negatively affect the social dynamic of clinical practice. As these aspects are difficult for RCTs to capture, accompanying qualitative studies may be called for.

5. Conclusion

In this article, we have done three things: we have analysed the rationale for medical AI RCTs and provided an overview over existing medical AI RCTs. On this basis, we considered different methodological challenges for AI RCTs, while pointing out ways to meet these challenges.

The concerns and recommendations that we have made above are by no means decisive or exhaustive. Instead, our article is meant to stimulate methodological reasoning for RCT trials testing AI assistance. An issue we have not discussed, for instance, concerns the validation of AI systems that are not *frozen* after the training phase (cf. Topol 2020). While this may not be relevant for current vision-based diagnostic systems, such AI systems could become crucial for personalized monitoring and treatment selection. The threat is that the AI system treats different patient demographics fairly during validation (according to some

[7] We thank an anonymous reviewer for this helpful comment.

fairness metric), but then develops a novel bias, resulting in unfair treatment. Building fair AI systems is difficult, precisely because there is no value-neutral way to select the training data, the objective function, the model, the benchmark task, the appropriate notion of fairness, and so on (Biddle, forthcoming; Johnson 2020). That difficulty is compounded by the fact that the algorithms continue to evolve after they are implemented. It may be extremely difficult to detect and overcome the bias once the system is launched into a clinical environment. The latest protocols are a welcome and important development. But protocols should also be developed that are tailored to the kinds of methodological issues that arise in these novel kinds of trials. Merely adapting the form of an RCT is no substitute for careful methodological reasoning. We hope that our discussion will stimulate the interest and attention of clinical methodologists and philosophers of medicine.

Acknowledgments
Both authors are supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence "Machine Learning—New Perspectives for Science", EXC 2064, project number 390727645).

References

Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. "A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy." In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. New York: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376718.

Berner, Eta S. and Tonya J. La Lande. 2016. "Overview of Clinical Decision Support Systems." In *Clinical Decision Support Systems: Theory and Practice*, edited by Eta Berner, 1–17. Health Informatics series. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-31913-1_1.

Biddle, Justin E. Forthcoming. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy*.

Bjerring, Jens Christian and Jacob Busch. 2020. "Artificial Intelligence and Patient-Centered Decision-Making." *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00391-6.

Cai, Carrie J., Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg et al. 2019. "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making." In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. New York: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300234.

Cartwright, Nancy. 2007. "Are RCTs the Gold Standard?" *BioSocieties* 2 (1): 11–20. https://doi.org/10.1017/S1745855207005029.

Cruz Rivera, Samantha, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert and the SPIRIT-AI and CONSORT-AI Working Group. 2020. "Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension." *The Lancet Digital Health* 2, no. 10: e549–e560. https://doi.org/10.1016/S2589-7500(20)30219-3.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210: 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

Erasmus, Adrian, Tyler Brunet and Eyal Fisher. 2020. "What is Interpretability?" *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00435-2.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau and Sebastian Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–118. https://doi.org/10.1038/nature21056.

Friedman, Lawrence, Curt D. Furberg and David L. DeMets. 2010. *Fundamentals of Clinical Trials*. Fourth edition. New York: Springer.

Fuller, Jonathan. 2019. "The Confounding Question of Confounding Causes in Randomized Trials." *British Journal for the Philosophy of Science* 70 (3): 901–926. https://doi.org/10.1093/bjps/axx015.

Gong, Dexin, Lianlian Wu, Jun Zhang, Ganggang Mu, Lei Shen, Jun Liu, Zhengqiang Wang et al. 2020. "Detection of Colorectal Adenomas with a Real-Time Computer-Aided System (ENDOANGEL): A Randomised Controlled Study." *The Lancet Gastroenterology & Hepatology* 5, no. 4: 352–361. https://doi.org/10.1016/S2468-1253(19)30413-3.

Grote, Thomas and Philipp Berens. 2020. "On the Ethics of Algorithmic Decision-Making in Healthcare." *Journal of Medical Ethics* 46, no. 3: 205–211. http://dx.doi.org/10.1136/medethics-2019-105586.

Grote, Thomas and Philipp Berens. Forthcoming. "Uncertainty, Evidence, and the Integration of Machine Learning into Medical Practice." *The Journal of Medicine and Philosophy*.

Gulshan, Varun, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan et al. 2016. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA* 316, no. 22: 2402–2410. https://doi.org/10.1001/jama.2016.17216.

Hernán, Miguel Angel. 2004. "A Definition of Causal Effect for Epidemiological Research." *Journal of Epidemiology & Community Health* 58, no. 4: 265–271. http://dx.doi.org/10.1136/jech.2002.006361.

Johnson, Gabbrielle M. 2020. "Algorithmic Bias: On the Implicit Biases of Social Technology." *Synthese*. https://doi.org/10.1007/s11229-020-02696-y.

Lalumera, Elisabetta and Stefano Fanti. 2019. "Randomized Controlled Trials for Medical Imaging: Conceptual and Practical Problems." *Topoi* 38, no. 2: 395–400. https://doi.org/10.1007/s11245-017-9535-z.

Lin, Haotian, Ruiyang Li, Zhenzhen Liu, Jingjing Chen, Yahan Yang, Hui Chen, Zhuoling Lin et al. 2019. "Diagnostic Efficacy and Therapeutic Decision-Making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial." *EClinicalMedicine* 9: 52–59. https://doi.org/10.1016/j.eclinm.2019.03.001.

Liu, Xiaoxuan, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston and and the SPIRIT-AI and CONSORT-AI Working Group. 2020. "Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension." *The Lancet Digital Health* 2, no. 10: e537–e548. https://doi.org/10.1016/S2589-7500(20)30218-1.

Liu, Xiaoxuan, Livia Faes, Aditya U. Kale, Siegfried K. Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran et al. 2019. "A Comparison of Deep Learning Performance Against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and Meta-Analysis." *The Lancet Digital Health* 1, no. 6: e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2.

Liu, Yuan, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada et al. 2020. "A Deep Learning System for Differential Diagnosis of Skin Diseases." *Nature Medicine* 26, no. 6: 900–908. https://doi.org/10.1038/s41591-020-0842-3.

Long, Erping, Haotian Lin, Zhenzhen Liu, Xiaohang Wu, Liming Wang, Jiewei Jiang, Yingying An et al. 2017. "An Artificial Intelligence Platform for the Multihospital Collaborative Management of Congenital Cataracts." *Nature Biomedical Engineering* 1, no. 2. https://doi.org/10.1038/s41551-016-0024.

McKinney, Scott M., Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back et al. 2020. "International Evaluation of an AI System for Breast Cancer Screening." *Nature* 577: 89–94. https://doi.org/10.1038/s41586-019-1799-6.

Mongan, John, Linda Moy and Charles E. Kahn. 2020. "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers." *Radiology: Artificial Intelligence* 2, no. 2: e200029. https://doi.org/10.1148/ryai.2020200029.

Mori, Yuichi, Shin-ei Kudo and Masashi Misawa. 2020. "Can Artificial Intelligence Standardise Colonoscopy Quality?" *The Lancet Gastroenterology & Hepatology* 5, no. 4: 331–332. https://doi.org/10.1016/S2468-1253(19)30407-8.

Nagendran, Myura, Yang Chen, Christopher A. Lovejoy, Anthony C. Gordon, Matthieu Komorowski, Hugh Harvey, Eric J. Topol, John P.A. Ioannidis, Gary S. Collins and Mahiben Maruthappu. 2020. "Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies." *BMJ* 368:m689. https://doi.org/10.1136/bmj.m689.

Oren, Ohad, Bernard J. Gersh and Deepak L. Bhatt. 2020. "Artificial Intelligence in Medical Imaging: Switching from Radiographic Pathological Data to Clinically Meaningful Endpoints." *The Lancet Digital Health* 2, no. 9: e486–e488. https://doi.org/10.1016/S2589-7500(20)30160-6.

Park, Yoonyoung, Gretchen Purcell Jackson, Morgan A. Foreman, Daniel Gruen, Jianying Hu and Amar K. Das. 2020. "Evaluating Artificial Intelligence in Medicine: Phases of Clinical Research." *JAMIA Open* 3, no. 3: 326–331. https://doi.org/10.1093/jamiaopen/ooaa033.

Russo, Federica and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21, no. 2: 157–170. https://doi.org/10.1080/02698590701498084.

Schaffner, Ken, ed. 1985. *Logic of Discovery and Diagnosis in Medicine*. Pittsburgh Series in Philosophy and History of Science. Berkeley: University of California Press.

Senn, Stephen. 2013. "Seven Myths of Randomization in Clinical Trials." *Statistics in Medicine* 32, no. 9: 1439–1450. https://doi.org/10.1002/sim.5713.

Steel, Daniel. 2011. "Causal Inference and Medical Experiments." Gifford, Fred (Ed.): *Handbook of the Philosophy of Science: Philosophy of Medicine*. Vol. 16. North-Holland: 159-185. https://doi.org/10.1016/B978-0-444-51787-6.50006-4.

Su, Jing-Ran, Zhen Li, Xue-Jun Shao, Chao-Ran Ji, Rui Ji, Ru-Chen Zhou, Guang-Chao Li et al. 2020. "Impact of a Real-Time Automatic Quality Control System on Colorectal Polyp and Adenoma Detection: A Prospective Randomized Controlled Study (With Videos)." *Gastrointestinal Endoscopy* 91, no. 2: 415–424. https://doi.org/10.1016/j.gie.2019.08.026.

Sullivan, Emily. Forthcoming. "Understanding from Machine Learning Models." *British Journal for the Philosophy of Science*.

Topol, Eric J. 2019. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine* 25, no. 1: 44–56. https://doi.org/10.1038/s41591-018-0300-7.

———. 2020. "Welcoming New Guidelines for AI Clinical Research." *Nature Medicine* 26, no. 9: 1318–1320. https://doi.org/10.1038/s41591-020-1042-x.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda et al. 2020. "Human-Computer Collaboration for Skin Cancer Recognition." *Nature Medicine* 26, no. 8: 1229–1234. https://doi.org/10.1038/s41591-020-0942-0.

Urbach, Peter. 1985. "Randomization and the Design of Experiments." *Philosophy of Science* 52, no. 2: 256–273. https://doi.org/10.1086/289243.

———. 1993. "The Value of Randomization and Control in Clinical Trials." *Statistics in Medicine* 12, no. 15–16: 1421–1431. https://doi.org/10.1002/sim.4780121508.

Varghese, Julian, Maren Kleine, Sophia Isabella Gessner, Sarah Sandmann and Martin Dugas. 2018. "Effects of Computerized Decision Support System Implementations on Patient Outcomes in Inpatient Care: A Systematic Review." *J Am Med Inform Assoc* 25, no. 5: 593–602. https://doi.org/10.1093/jamia/ocx100.

Vleugels, Jasper L.A., Yark Hazewinkel, Paul Fockens and Evelien Dekker. 2017. "Natural History of Diminutive and Small Colorectal Polyps: A Systematic Literature Review." *Gastrointestinal Endoscopy* 85, no. 6 (June): 1169–1176. https://doi.org/10.1016/j.gie.2016.12.014.

Wang, Pu, Tyler M. Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu et al. 2019. "Real-Time Automatic Detection System Increases Colonoscopic Polyp and Adenoma Detection Rates: A Prospective Randomised Controlled Study." *Gut* 68, no. 10: 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500.

Wang, Pu, Xiaogang Liu, Tyler M. Berzin, Jeremy R. Glissen Brown, Peixi Liu, Chao Zhou, M.M. Lei Lei et al. 2020. "Effect of a Deep-Learning Computer-Aided Detection System on Adenoma Detection During Colonoscopy (CADe-DB Trial): A Double-Blind Randomised Study." *The Lancet Gastroenterology & Hepatology* 5, no. 4: 343–351. https://doi.org/10.1016/S2468-1253(19)30411-X.

Wijnberge, Marije, Bart F. Geerts, Liselotte Hol, Nikki Lemmers, Marijn P. Mulder, Patrick Berge, Jimmy Schenk et al. 2020. "Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension Vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial." *JAMA* 323, no. 11: 1052–1060. https://doi.org/10.1001/jama.2020.0592.

Worrall, John. 2002 "*What* Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69, no. 3: 316–30. https://doi.org/10.1086/341855.

———. 2007. "Why There's No Cause to Randomize." *British Journal for the Philosophy of Science* 58, no. 3: 451–488. https://doi.org/10.1093/bjps/axm024.

———. 2010. "Evidence: Philosophy of Science Meets Medicine." *Journal of Evaluation in Clinical Practice* 16, no. 2: 356–362. https://doi.org/10.1111/j.1365-2753.2010.01400.x.

Wu, Lianlian, Jun Zhang, Wei Zhou, Ping An, Lei Shen, Jun Liu, Xiaoda Jiang et al. 2019. "Randomised Controlled Trial of WISENSE, a Real-Time Quality Improving System for Monitoring

Blind Spots During Esophagogastroduodenoscopy." *Gut* 68, no. 12: 2161–2169. https://doi.org/10.1136/gutjnl-2018-317366.