

BMJ Open Aversion to pragmatic randomised controlled trials: three survey experiments with clinicians and laypeople in the USA

Randi L Vogt ¹, Patrick R Heck ¹, Rebecca M Mestechkin ¹,
Pedram Heydari ², Christopher F Chabris ¹, Michelle N Meyer ¹

To cite: Vogt RL, Heck PR, Mestechkin RM, *et al.* Aversion to pragmatic randomised controlled trials: three survey experiments with clinicians and laypeople in the USA. *BMJ Open* 2024;**14**:e084699. doi:10.1136/bmjopen-2024-084699

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2024-084699>).

RLV and PRH contributed equally.
CFC and MNM contributed equally.

RLV and PRH are joint first authors.
CFC and MNM are joint senior authors.

Received 26 January 2024
Accepted 31 July 2024



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Bioethics & Decision Sciences, Geisinger, Danville, Pennsylvania, USA

²Economics, Northeastern University—Boston Campus, Boston, Massachusetts, USA

Correspondence to

Dr Michelle N Meyer;
michellenmeyer@gmail.com

ABSTRACT

Objectives Pragmatic randomised controlled trials (pRCTs) are essential for determining the real-world safety and effectiveness of healthcare interventions. However, both laypeople and clinicians often demonstrate experiment aversion: preferring to implement either of two interventions for everyone rather than comparing them to determine which is best. We studied whether clinician and layperson views of pRCTs for COVID-19, as well as non-COVID-19, interventions became more positive during the pandemic, which increased both the urgency and public discussion of pRCTs.

Design Randomised survey experiments.

Setting Geisinger, a network of hospitals and clinics in central and northeastern Pennsylvania, USA; Amazon Mechanical Turk, a research participant platform used to recruit online participants residing across the USA. Data were collected between August 2020 and February 2021.

Participants 2149 clinicians (the types of people who conduct or make decisions about conducting pRCTs) and 2909 laypeople (the types of people who are included in pRCTs as patients). The clinician sample was primarily female (81%), comprised doctors (15%), physician assistants (9%), registered nurses (54%) and other medical professionals, including other nurses, genetic counsellors and medical students (23%), and the majority of clinicians (62%) had more than 10 years of experience. The layperson sample ranges in age from 18 to 88 years old (mean=38, SD=13) and the majority were white (75%) and female (56%).

Outcome measures Participants read vignettes in which a hypothetical decision-maker who sought to improve health could choose to implement intervention A for all, implement intervention B for all, or experimentally compare A and B and implement the superior intervention. Participants rated and ranked the appropriateness of each decision. Experiment aversion was defined as the degree to which a participant rated the experiment below their lowest-rated intervention.

Results In a survey of laypeople administered during the pandemic, we found significant aversion to experiments involving catheterisation checklists and hypertension drugs unrelated to the treatment of COVID-19 (Cohen's $d=0.25-0.46$, $p<0.001$). Similarly, among both laypeople and clinicians, we found significant aversion to most (comparing different checklist, pronging and mask

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The decision-science approach used in this paper enables measurement of aversion towards pragmatic randomised controlled trials (pRCTs) in large and diverse samples of clinicians and laypeople.
- ⇒ The size of the experiment aversion effect is measured in eight vignettes in the layperson sample and four vignettes in the clinician sample that describe a range of pRCTs, from pharmaceutical and non-pharmaceutical medical interventions to public health interventions, some of which are specific to the COVID-19 pandemic and others of which describe more general medical situations.
- ⇒ The large sample sizes ensured sufficient statistical power to detect experiment aversion in each vignette and sample.
- ⇒ The samples may not perfectly represent all healthcare professionals or members of the general public as they are convenience samples of clinicians at a specific hospital system in the USA and laypeople on a specific online crowdworking platform.
- ⇒ Participants expressed attitudes and judgements about the appropriateness of carrying out pRCTs or implementing policies but were not in a position to make a real decision to execute the pRCTs or policies.

protocols; Cohen's $d=0.17-0.56$, $p<0.001$) but not all (comparing school reopening protocols; Cohen's $d=0.03$, $p=0.64$) non-pharmaceutical COVID-19 experiments. Interestingly, we found the lowest experiment aversion to pharmaceutical COVID-19 experiments (comparing new drugs and new vaccine protocols for treating the novel coronavirus; Cohen's $d=0.04-0.12$, $p=0.12-0.55$). Across all vignettes and samples, 28%–57% of participants expressed experiment aversion, whereas only 6%–35% expressed experiment appreciation by rating the trial higher than their highest-rated intervention.

Conclusions Advancing evidence-based medicine through pRCTs will require anticipating and addressing experiment aversion among patients and healthcare professionals.

Study registration <http://osf.io/6p5c7/>.

INTRODUCTION

Pragmatic randomised controlled trials (pRCTs) are crucial for understanding how to safely, effectively, and equitably prevent and treat disease and deliver healthcare. Randomised evaluation is the gold standard in medicine, largely because it permits one to infer that an intervention *caused* an outcome, such as reduction of symptoms or improvement in a biomarker. Randomised experiments have repeatedly upended conventional clinical wisdom and the results of observational studies^{1,2} and are urgently needed to evaluate new technologies.^{3,4} Compared with more explanatory trials, trials that are further towards the pragmatic end of the spectrum⁵ evaluate effectiveness of the intervention in more real-world contexts. Such pragmatism is critical for ensuring that causal evidence from randomised evaluation speaks to the effects of interventions in the circumstances in which they would be implemented (or maintained).

Yet despite their importance to healthcare quality and safety, pRCTs often prove controversial—even when they compare interventions that are within the standard of care or are otherwise unobjectionable, and about which the relevant expert community is in equipoise. Several recently published pRCTs—including the Surfactant, Positive Pressure, and Oxygenation Randomized Trial (SUPPORT),⁶ the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) trial,⁷ and the Individualized Comparative Effectiveness of Models Optimizing Patient Safety and Resident Education (iCOMPARE) trial⁸—have received considerable criticism from physician-scientists, ethicists and regulators^{9,10} and in the public square.^{11–14} Although criticisms of pRCTs can be complex, nuanced, and sometimes valid, many appear to reflect a rejection of the very idea that a randomised experiment was conducted as opposed to simply giving everyone one of the interventions that was trialled. Our research applies concepts and methods from the behavioural and decision sciences to systematically explore whether, when, and why people might genuinely object to running pRCTs in healthcare, public health, and other domains.

In prior studies—inspired by several ‘notorious pRCTs’, including technology industry ‘A/B tests’^{15–17}—we confirmed that substantial shares of both laypeople and clinicians can be averse to randomised evaluation of efforts to improve health.^{18,19} People rated a pRCT designed to compare the effectiveness of two interventions as less appropriate than the average appropriateness of implementing either one, untested, for everyone. We called this phenomenon the ‘A/B effect’.¹⁸ In some cases, the lower average rating of an experiment could be driven not by dislike of experiments, per se, but by many raters’ belief that one of the experiment’s arms is inferior to the other.^{18–21} Importantly, such beliefs are often based on intuition rather than evidence and have the potential to undermine evidence-based medicine. Yet this form of experiment rejection is not illogical, given the individual’s own beliefs. We also, however, documented a more peculiar (if no less dangerous) phenomenon of ‘experiment

aversion’, which occurred when people rated the pRCT as significantly less appropriate than implementing *their own* least-preferred intervention contained within the trial. In this pattern of decision-making, in other words, people who perceive that one intervention is good and the other is less good prefer that everyone receive the less good (or even bad) intervention rather than half the people receiving the better one, and without comparing the two to determine whether one is really better than the other.¹⁹ Such judgments could reflect a more general scepticism about or opposition to pRCTs, at least within specific domains of inquiry. For instance, people may be averse to the inequality or disparate treatment that is necessarily (temporarily) imposed by any RCT (pRCT or otherwise), the uncertainty signalled by agents (often trusted experts) who decide they do not already know what works and need to conduct a pRCT, the process of assigning people to treatments ‘randomly’ as opposed to using expert judgement, or something else viewed as undesirable. Both patterns of negative sentiments about experiments (the ‘A/B effect’ and ‘experiment aversion’) can impede efforts to assure and improve health outcomes.

The COVID-19 pandemic presented the potential for an inflection point in attitudes towards pRCTs. In April 2020, 72 COVID-19 drug trials were already underway²² and more traditional, explanatory RCTs became daily, front-page news. Because explanatory and pragmatic RCTs share many key features that participants in our prior research often cited as partial explanations for their lower ratings of experiments—including random assignment to different conditions¹⁸—the sustained exposure to explanatory RCTs during the pandemic might have educated people about the value of healthcare pRCTs, too, and/or made them seem less exceptional and more normative. Our previous research also suggests that another cause of experiment aversion is an illusion of knowledge—a (mis)perception that experts already must know what works best and should simply implement those interventions without further study. But COVID-19 was a novel disease, and—at least in the case of pharmaceutical interventions—no sensible person thought the correct treatments were already obvious. People, therefore, may have been less averse to COVID-19 pRCTs (eg, trials comparing COVID-19 proning protocols or masking rules) than to pRCTs that test interventions for familiar conditions or problems, such as hypertension or hospital-acquired infections. On the other hand, because of the urgency attached to COVID-19, people may have been *more* averse to COVID-19 RCTs, being even less inclined to risk giving someone a treatment that might turn out to ‘lose’ in a comparison study.^{23,24} Finally, even if the pandemic did not affect public attitudes towards explanatory or pragmatic RCTs, it could have affected the attitudes of clinicians, many of whom were involved in COVID-19 research. Because clinicians strongly influence whether particular RCTs (both explanatory and pragmatic) are conducted, their attitudes matter. Here,

we investigated layperson and clinician attitudes towards pRCTs in the first year of the pandemic by conducting a series of preregistered studies between August 2020 and February 2021.

METHODS

Study setting

The study was conducted online using the Qualtrics platform.²⁵ For the layperson sample, we used the CloudResearch service^{26 27} to recruit adult crowd workers on Amazon Mechanical Turk²⁸ living in the USA to participate in a brief online survey. These services provide samples that are broadly representative of the US population and are well accepted in social science research as providing as good or better quality, diverse samples of research participants than common convenience samples such as student volunteers, with results that are similar to probability sampling methods.^{29–31} Clinicians of various levels in healthcare were recruited by email (following a procedure successfully used previously¹⁸) from Geisinger, a network of hospitals and clinics in central and north-eastern Pennsylvania, USA with a medical school and a research institute. Geisinger's IRB determined that these surveys were exempted (IRB# 2017-0449).

Study design

Data were collected between August 2020 and February 2021 (online supplemental table S1). First, we used decision-making vignettes from our previous work to ask whether the extraordinary publicity around (primarily explanatory) COVID-19 RCTs reduced general healthcare experiment aversion by the public. Next, we adapted these vignettes to determine whether the public was averse to pRCTs on pharmaceutical and/or non-pharmaceutical interventions (NPIs) for COVID-19. Finally, we recruited a large clinician sample to investigate how their attitudes compared with those of laypeople.

Participants were evenly randomised to read one of the vignettes. Randomisation was accomplished using a proprietary least filled quota algorithm built into the Qualtrics survey software, such that aside from participants who withdrew before completing the survey, the same number of participants are allocated to each vignette (see online supplemental materials for additional details). Each vignette described a problem that the decision-maker could address in one of three ways: by implementing intervention A for all patients or relevant members of the public (A); by implementing intervention B for all patients or relevant members of the public (B); or by conducting an experiment in which patients or relevant members of the public are randomly assigned to A or B and the superior intervention is then implemented for all (A/B test). For example, in Best Anti-Hypertensive Drug, some doctors in a walk-in clinic prescribe 'Drug A' while others prescribe 'Drug B' (both of which are affordable, tolerable and FDA approved), and 'Dr Jones' prescribes either A for all his hypertensive patients, B for

all those patients or runs a randomised experiment to compare the effectiveness of A and B. (See table 1 for two additional vignette examples, online supplemental table S2 for all vignette names, and pp8–13 in the online supplemental materials for all vignette text.) To develop the vignettes, we consulted the literature and our knowledge, as experts in bioethics and psychological science, of pRCTs that have historically proved controversial (see online supplemental table S3 for motivations for all vignettes). All vignettes describe an RCT that is highly pragmatic in nature (ie, high on PRECIS-2 eligibility, recruitment, setting, organisation, follow-up and primary outcome domains⁵). For instance, all patients with the relevant condition who attend the clinic/hospital for care become members of the trial and the trial is situated within the clinic/hospital where their care would typically take place. (Similarly, in the public health scenarios, all students in the school district and all residents of the state where these trials occur are included in the trial.) In addition, our vignettes are silent about whether consent will be obtained. Trials that include only those who opt into them are less pragmatic if they are testing the effectiveness of an intervention that would be imposed on people as a matter of policy or practice. IRBs customarily waive consent when it would make low-risk pRCTs impracticable, including by rendering the results uninformative about how an intervention would fare in practice.³² In separate work, we found that substantial shares of people object to such experiments even when we specify that consent will be obtained.³³

Next, following a standard decision-science approach commonly used in social and moral psychology for evaluating decisions,³⁴ participants rated each option on a scale of appropriateness from 1 ('very inappropriate') to 5 ('very appropriate'), with 3 as a neutral midpoint. Participants then rank ordered the options from best to worst and provided demographic information.

Participants

Based on a power analysis, we determined that recruiting ~350 participants (laypeople and clinicians) per vignette (Catheterization Safety Checklist, Best Anti-Hypertensive Drug, Intubation Safety Checklist, Best Corticosteroid Drug, Masking Rules, School Reopening, and Ventilator Proning) would yield 95% power to detect an effect as small as Cohen's $d=0.19$ at $\alpha=0.05$. These sample sizes are consistent with our previous work using the same methods (but different vignettes¹⁹).

For Best Vaccine, based on a prior study (see online supplemental materials for full details), we hypothesised a smaller effect size, which resulted in a power analysis that determined that recruiting ~450 lay participants (a sample size consistent with our previous work¹⁹) would yield 80% power to detect an effect as small as Cohen's $d=0.13$ and 95% power to detect as small as Cohen's $d=0.17$. For the clinician sample, we based our power analysis for Best Vaccine on the number of responses we collected in the first clinician survey testing the Masking

Table 1 Vignette text for Catheterization Safety Checklist and Ventilator Proning

	Catheterization Safety Checklist	Ventilator Proning
Background	Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections.	Some coronavirus (COVID-19) patients have to be sedated and placed on a ventilator to help them breathe. Even with a ventilator, these patients can have dangerously low blood oxygenation levels, which can result in death. Current standards suggest that laying ventilated patients on their stomach for 12–16 hours per day can reduce pressure on the lungs and might increase blood oxygen levels and improve survival rates.
Intervention A	A hospital director wants to reduce these infections, so he decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure printed on the back. All patients having this procedure will then be treated by doctors with this list attached to their clothing.	A hospital director wants to save as many ventilated COVID-19 patients as possible, so he decides that all of these patients will be placed on their stomach for 12–13 hours per day.
Intervention B	A hospital director wants to reduce these infections, so he decides to hang a poster with a list of standard safety precautions for this procedure in all procedure rooms. All patients having this procedure will then be treated in rooms with this list posted on the wall.	A hospital director wants to save as many ventilated COVID-19 patients as possible, so he decides that all of these patients will be placed on their stomach for 15–16 hours per day.
A/B test	A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall. After a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.	A hospital director thinks of two different ways to save as many ventilated COVID-19 patients as possible, so he decides to run an experiment by randomly assigning ventilated COVID-19 patients to one of two test conditions. Half of these patients will be placed on their stomach for 12–13 hours per day. The other half of these patients will be placed on their stomach for 15–16 hours per day. After one month, the director will have all ventilated COVID-19 patients treated in whichever way turns out to have the highest survival rate.

Rules, Intubation Safety Checklist and Best Corticosteroid vignettes. We assumed ~900 responses, which we determined would yield 95% power to detect an effect as small as $d=0.12$.

Across all vignettes, there were a total of 2909 lay participants. They ranged in age from 18 to 88 with a mean age of 38 years old ($SD=12.8$). The majority of participants were white (75%), female (56%) and college educated (30% having completed some college, 36% having earned a 4-year degree and 21% having earned a graduate degree; 21% of participants had a Science, Technology, Engineering, or Math (STEM) degree) with a median household income of \$40 000 to \$60 000. The sample is more liberal (44%) and Democrat (38%) than conservative (28%) and Republican (21%) and a plurality of participants identified as non-religious (38%).

The clinician sample ($N=2149$) was comprised of doctors (15%), physician assistants (9%), nurse practitioners (5%), nurses (67%; registered nurses: 54%; licensed practical nurses: 12%; other: 1%) and other medical professionals (including genetic counsellors and medical students; 4%). We determined the ratio of different types of clinicians from their self-reported position in the survey. We did not estimate in advance the proportion of certain types of clinicians who would respond. The majority of the clinicians were female (81%) and had been working in healthcare for more than 10 years (62%). A majority of clinicians reported being somewhat or moderately comfortable with research methods

and statistics (77%) and had two sources of formal or informal training or education in research methods and statistics (eg, undergraduate, professional school or post-graduate coursework; 58%). (In these clinician samples, because survey responses were made fully anonymous to encourage greater participation and honest responding, we were unable to restrict participation in later waves to clinicians who had not participated in earlier waves. Therefore, some clinicians who completed the Best Vaccine vignette may have earlier completed the Masking Rules, Intubation Safety Checklist, and Best Corticosteroid Drug vignettes.) See online supplemental tables S4–S5 for detailed demographics of lay participants and clinicians by vignette.

Data analysis

We define the ‘A/B effect’ as the degree to which participants’ ratings of the A/B test were lower than the average of their ratings of implementing A and B.¹⁸ ‘Experiment aversion’ is the degree to which participants rated the A/B test lower than their own lowest-rated intervention (either A or B for each person).¹⁹ ‘Experiment appreciation’ is the opposite: the degree to which the experiment is rated higher than each participant’s highest-rated intervention. For all measures, we performed paired t -tests at $\alpha=0.05$ and calculated Cohen’s d recovered from the t -statistic, sample size, and correlation between the two measures being compared.^{35 36} We also calculated the percentage of participants who ranked the A/B test

as the worst (or best) option the decision-maker could implement as well as the percentage of participants who showed an A/B effect, were experiment averse, or were experiment appreciative. We analysed data using R version 4.3.0. Participant response data, preregistrations, materials and analysis code have been deposited at OSF.³⁷

Patient and public involvement

We included laypeople as participants in our studies because they are typically included in pRCTs as patients or (in the case of some public health pRCTs and pRCTs in other domains) as members of the public and are, therefore, important stakeholders. Decisions about whether to participate in or conduct pRCTs are made against the backdrop of individuals' personal views and/or anticipation of potential backlash or other public reactions; therefore, how patients and clinicians feel about experiments is relevant to if and how advancements in health-care are made. All participant responses were anonymous and, thus, results cannot be disseminated back to our participants.

RESULTS

In the following results, we group the vignettes by theme: those eliciting lay participants' sentiments about pRCTs unrelated to the treatment of COVID-19, those eliciting lay participants' sentiments about pRCTs related to the treatment of, prevention of or public health response to COVID-19, and those eliciting clinician sentiments about pRCTs related to the treatment of, prevention of or public health response to COVID-19.

Lay sentiments about pRCTs

To elicit lay sentiments about pRCTs, participants responded to one of two vignettes: Catheterization Safety Checklist (which described two locations where a hospital director could display a safety checklist for clinicians; see table 1; $n=343$) or Best Anti-Hypertensive Drug (which described two drugs a doctor could prescribe for his hypertensive patients; $n=357$).

We found substantial negative reactions to A/B testing in both vignettes (table 2), replicating our prepandemic findings.^{18 19} Although in most cases, the mean rating of the A/B test was near the neutral midpoint, implementing policies was substantially preferred to A/B testing (figure 1A) and large proportions of participants objected to the A/B test (figure 1B). In Catheterization Safety Checklist (figure 1A), we found evidence of the A/B Effect: participants rated the A/B test significantly below the average ratings they gave to implementing interventions A and B ($d=0.69$, 95% CI (0.53 to 0.85); online supplemental table S6A). Here, 41%±5% (95% CI) of participants expressed experiment aversion (rating the A/B test lower than their own lowest-rated intervention; $d=0.25$, 95% CI (0.11 to 0.39); online supplemental table S6A). When ranking the three options from best

to worst, only 32% placed the A/B test first, while 48% placed it last (online supplemental table S6A).

We also observed an A/B effect in Best Anti-Hypertensive Drug (figure 1A; $d=0.52$, 95% CI (0.36 to 0.68); online supplemental table S6A), where 44%±5% also expressed experiment aversion ($d=0.46$, 95% CI (0.30 to 0.52); online supplemental table S6A). Notably, participants were averse to this experiment even though there is no reason to prefer 'Drug A' to 'Drug B', and patients are effectively already randomised to A or B based on which clinician happens to see them—which occurs wherever unwarranted variation in practice determines treatments, such as walk-in clinics and emergency departments. Here, however, similar proportions of people ranked the A/B test best and worst (50% vs 45%; $p=0.16$; online supplemental table S6A).

These levels of experiment aversion near the height of the pandemic were slightly (but not significantly) higher than those we observed among similar laypeople in 2019 (41%±5% in 2020 vs 37%±6% in 2019 for Catheterization Safety Checklist, $p=0.31$; 44%±5% in 2020 vs 40%±6% in 2019 for Best Anti-Hypertensive Drug, $p=0.32$).¹⁹

Lay sentiments about COVID-19 pRCTs

To elicit lay sentiments about COVID-19 pRCTs, we asked lay participants to read one of the following vignettes: Masking Rules (which described two masking policies, of varying scope; $n=360$); School Reopening (which described two school schedules designed to increase social distancing; $n=339$); Best Vaccine (which described two types of vaccine—mRNA vs inactivated virus; $n=450$); Ventilator Proning (which described two protocols for positioning ventilated patients with COVID-19; see table 1; $n=357$); Intubation Safety Checklist (adapted from Catheterization Safety Checklist above to apply to COVID-19; $n=347$) and Best Corticosteroid Drug (adapted from Best Anti-Hypertensive Drug above to apply to COVID-19; $n=357$).

In all six COVID-19 vignettes, we found evidence of the A/B effect (table 2 and figure 2A). In three, however, we did not find experiment aversion: Best Vaccine^[1], Best Corticosteroid Drug and School Reopening. In the first two of these, participants rated the two interventions very similarly and the experiment only slightly lower (figure 2B). These vignettes also elicited the largest proportion of participants (65% in Best Vaccine and 56% in Best Corticosteroid Drug; online supplemental table S6B) in any vignette who ranked the A/B test best among the three options, compared with 31%–34% of participants who ranked it worst (online supplemental table S6B). In School Reopening, experiment aversion was not observed because participants on average clearly preferred intervention B to A (figure 2B) and rated the experiment similar to intervention A (figure 2A).^{20 21}

ⁱSee online supplemental table S6D for results from a previous version of Best Vaccine, which unintentionally implied that vignette participants could choose their vaccine.

Table 2 Sentiments about experiments by vignette and population

	Negative sentiment				Positive sentiment			
	Experiment aversion	A/B effect	More people averse than appreciative?	More people rank A/B test worst than best?	More people rank A/B test best than worst?	More people appreciative than averse?	Reverse A/B effect	Experiment appreciation
Lay sentiments about pRCTs								
Catheterization Safety Checklist	✓	✓	✓	✓				
Best Anti-Hypertensive Drug	✓	✓	✓					
Lay sentiments about COVID-19 pRCTs								
Ventilator Proning	✓	✓	✓					
School Reopening		✓	✓	✓				
Masking Rules	✓	✓	✓	✓				
Intubation Safety Checklist	✓	✓	✓	✓				
Best Corticosteroid Drug		✓			✓			
Best Vaccine		✓			✓			
Clinician sentiments about COVID-19 pRCTs								
Masking Rules	✓	✓	✓	✓				
Intubation Safety Checklist	✓	✓	✓	✓				
Best Corticosteroid Drug	✓	✓	✓					
Best Vaccine		**			✓			

Experiment aversion refers to the difference between the lowest-rated intervention and the rating of the A/B test. The A/B effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. The Reverse A/B effect refers to the difference between the rating of the A/B test and the average rating of the two interventions. Experiment appreciation refers to the difference between the rating of the A/B test and the rating of the highest-rated intervention. See online supplemental table S6A–C for detailed results (including Cohen's ds and 95% CIs) for all measures of sentiment about experiments. Checkmarks (✓) represent a statistically significant effect at $p < 0.05$. In one case, the checkmark is followed by an asterisk (*). This indicates that while the effect reaches statistical significance, the effect size is very small and might have only reached significance due to the large sample size (three times as large as that for other vignettes). Variables to the right of the vertical line are the reverse of those on the left. If no checkmark appears in either of the corresponding columns to the left and right of the vertical line (eg, 'More people rank A/B test worst than best?' and 'More people rank A/B test best than worst?'), that means that there is no statistically significant difference between the proportion of people ranked A/B test worst and the proportion of people who ranked the A/B test best).

pRCTs, pragmatic randomised controlled trials.

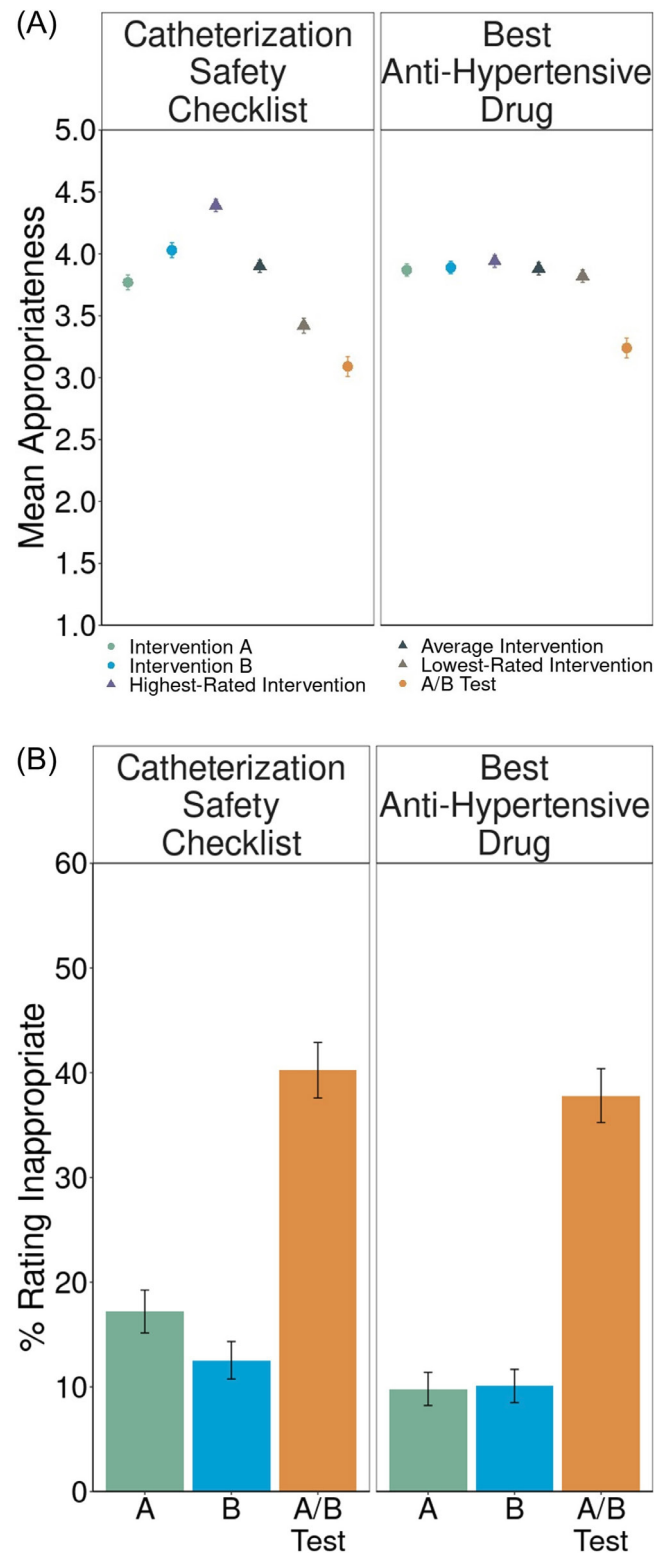


Figure 1 Lay sentiments about pragmatic randomised controlled trials (pRCTs). (A) Mean appropriateness ratings, on a 1–5 scale, with standard errors, for intervention A, intervention B, the highest-rated intervention, the average intervention, the lowest-rated intervention and the A/B test. Circles represent measures directly collected from participants. Triangles represent averages derived from the direct measures. The distance of the mean appropriateness of the lowest-rated intervention (brown triangle) minus the mean appropriateness of the A/B test (orange circle) represents experiment aversion. The distance of the mean appropriateness of the average intervention (gray triangle) minus the mean appropriateness of the A/B test (orange circle) represents the A/B effect. The distance of the mean appropriateness of the A/B test (orange circle) minus the mean appropriateness of the highest-rated intervention (purple triangle) represents experiment appreciation. (B) Appropriateness ratings transformed into percentages and standard errors of participants objecting (defined as assigning a rating of 1 or 2—‘very inappropriate’ or ‘somewhat inappropriate’—on a 1–5 scale) to implementing intervention A, intervention B and the A/B test.

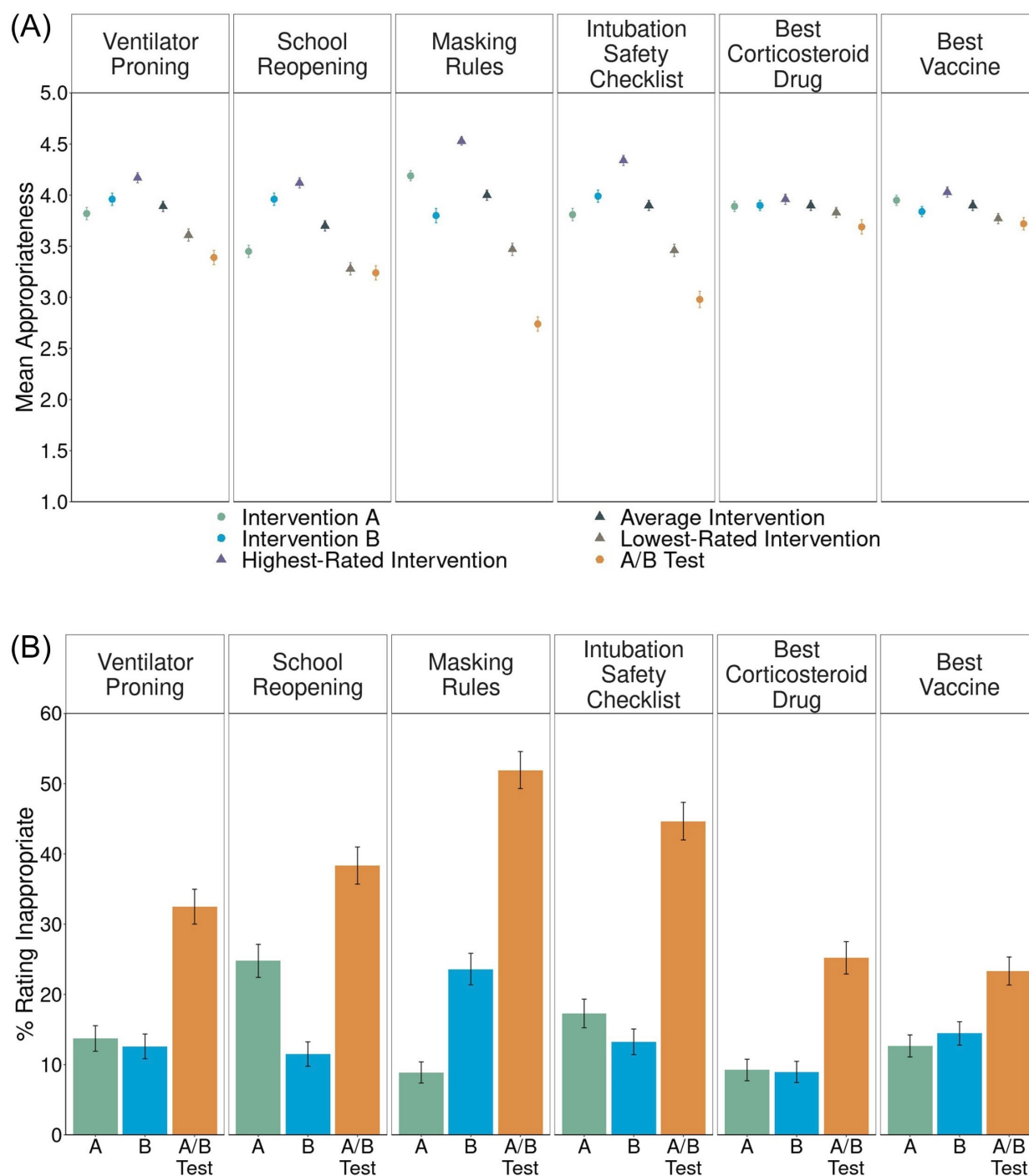


Figure 2 Lay sentiments about COVID-19 pragmatic randomised controlled trials (pRCTs). (A) Mean appropriateness ratings, on a 1–5 scale, with standard errors, for intervention A, intervention B, the highest-rated intervention, the average intervention, the lowest-rated intervention and the A/B test. Circles represent measures directly collected from participants. Triangles represent averages derived from the direct measures. The distance of the mean appropriateness of the lowest-rated intervention (brown triangle) minus the mean appropriateness of the A/B test (orange circle) represents experiment aversion. The distance of the mean appropriateness of the average intervention (gray triangle) minus the mean appropriateness of the A/B test (orange circle) represents the A/B effect. The distance of the mean appropriateness of the A/B test (orange circle) minus the mean appropriateness of the highest-rated intervention (purple triangle) represents experiment appreciation. (B) Appropriateness ratings transformed into percentages and standard errors of participants objecting (defined as assigning a rating of 1 or 2—‘very inappropriate’ or ‘somewhat inappropriate’—on a 1–5 scale) to implementing intervention A, intervention B and the A/B test.

Fifty-three per cent of participants ranked intervention B as the best of the three options (compared with 17% choosing intervention A and 30% choosing the A/B test; online supplemental table S6B).

In the other three vignettes, participants rated the A/B test condition as significantly less appropriate than their lowest-rated intervention (Masking Rules: $d=0.56$, 95% CI (0.41 to 0.71); Ventilator Proning: $d=0.17$, 95% CI (0.04 to 0.30); Intubation Safety Checklist: $d=0.36$, 95% CI (0.21 to 0.49)). These levels of aversion to COVID-19 RCTs are similar to the levels of aversion to non-Covid-19 RCTs both before¹⁹ and during the pandemic (see above).

Clinician sentiments about COVID-19 pRCTs

Clinicians responded to one^[ii] of four COVID-19-related vignettes: Masking Rules ($n=349$), Intubation Safety Checklist ($n=271$), Best Corticosteroid Drug ($n=275$) or Best Vaccine ($n=1254$). We observed an A/B effect in all four vignettes (figure 3A,B). In two, clinicians, like laypeople, were also significantly experiment averse (Masking Rules: $d=0.74$, 95% CI (0.57 to 0.91); online supplemental table S6C; Intubation Safety Checklist: $d=0.30$, 95% CI (0.15 to 0.45); online supplemental table S6C). In Best Vaccine, clinicians, like laypeople, did not show any significant difference in their ratings of the A/B test and their lowest-rated intervention ($d=-0.03$, 95% CI (-0.10 to 0.04); online supplemental table S6C). Again, like laypeople, 58% of clinicians ranked the Best Vaccine A/B test as the best of the three options, the highest proportion of any clinician-rated vignette.

Clinicians differed from laypeople in their response to Best Corticosteroid Drug. Laypeople did not show experiment aversion, but clinicians rated the A/B test as significantly less appropriate than their lowest-rated intervention ($d=0.49$, 95% CI (0.32 to 0.66); online supplemental table S6C). This difference may be due to clinicians' greater familiarity with the treatment of COVID-19. Clinicians may also have seen an urgent need for any drugs to treat COVID-19²⁴ and thus rated adopting a clear treatment intervention as more appropriate than an RCT.

Heterogeneity in experiment aversion

Collapsed across studies, political ideology explained 1.5% of the variance ($p<0.001$) in sentiments about experiments, with conservatives slightly less averse to experiments than liberals. Less or no variation was explained by all other demographics, including educational attainment (0.2%, $p=0.008$), STEM degree (0.1%, $p=0.15$) and prescribers versus other clinicians (0.2%, $p=0.061$); see

online supplemental figure S8-11 and p28 in the Supplemental Materials for further discussion.

DISCUSSION

In three preregistered survey experiments, we observed considerable experiment aversion among laypeople during the first year of the COVID-19 pandemic, despite increased exposure to the nature and purpose of (largely explanatory) RCTs. Neither laypeople nor clinicians were overall less averse to COVID-19 pRCTs, despite the fact that confidence in anyone's knowledge of what works should have been even more circumscribed than in the everyday contexts of hypertension and catheter infections. To the contrary, most COVID-19 vignettes were met with experiment aversion. This is consistent with an emphasis during the pandemic that we must 'do' instead of 'learn', a false dichotomy that fails to recognise that implementing an untested intervention is itself a non-consensual experiment from which, unlike an RCT, little or nothing can be learnt.³⁸⁻⁴⁰ Participants may have been averse to the uncertainty that the decision to conduct an experiment conveys. They may have perceived the experiment as more risky than implementing either of the policies it contains. Or they may have experienced hindsight bias, believing that the experiment was unfair to whom ever received the least effective policy, neglecting the fact that the results were not known in advance. For whatever reason, across all vignettes and samples, between 28% and 57% of participants demonstrated experiment aversion, while only 6%–35% demonstrated experiment appreciation (by rating the pRCT higher than their highest-rated intervention).

Although in most cases, the mean rating of the A/B test was near the neutral midpoint, in none of our 12 studies were more people appreciative of than averse to the pRCT, in none was the average pRCT rating higher than the average intervention rating, and in none was the pRCT rating higher than each participant's highest-rated intervention, on average. Notably, unlike trials with placebo or no-contact controls, the A/B tests in our vignettes compared two active, plausible interventions, neither of which was obviously known *ex ante* to be superior. Yet substantial shares of participants still preferred that one intervention simply be implemented without bothering to determine which (if either) worked best.

The most positive sentiment towards experiments was observed in both laypeople and clinicians in the vignettes involving COVID-19 drugs and vaccines. Here, we observed the highest proportions of participants who demonstrated experiment appreciation (26%–35%) and who ranked the pRCT first (49%–65%). This result could be explained by differences in the pRCT length (ranging from 1 to 12 months) and perceived severity of the pRCT outcome ('best outcome' and 'fewest cases of COVID-19' in Best Corticosteroid and Best Vaccine, respectively, vs, eg, 'highest survival rate' in Ventilator Proning). This result is also consistent with our previous findings that

ⁱⁱ Clinicians in the first survey were randomly assigned to one of the three vignettes (Masking Rules, Intubation Safety Checklist and Best Corticosteroid Drug) and then completed the remaining vignettes in random order. For consistency with the rest of this project and with our previous approach,¹⁸ we analysed data from this survey as a between-subjects design where we only consider the first vignette that every participant completed. See online supplemental material table S7 and pp. 27–28 in the Online supplemental material for further discussion.

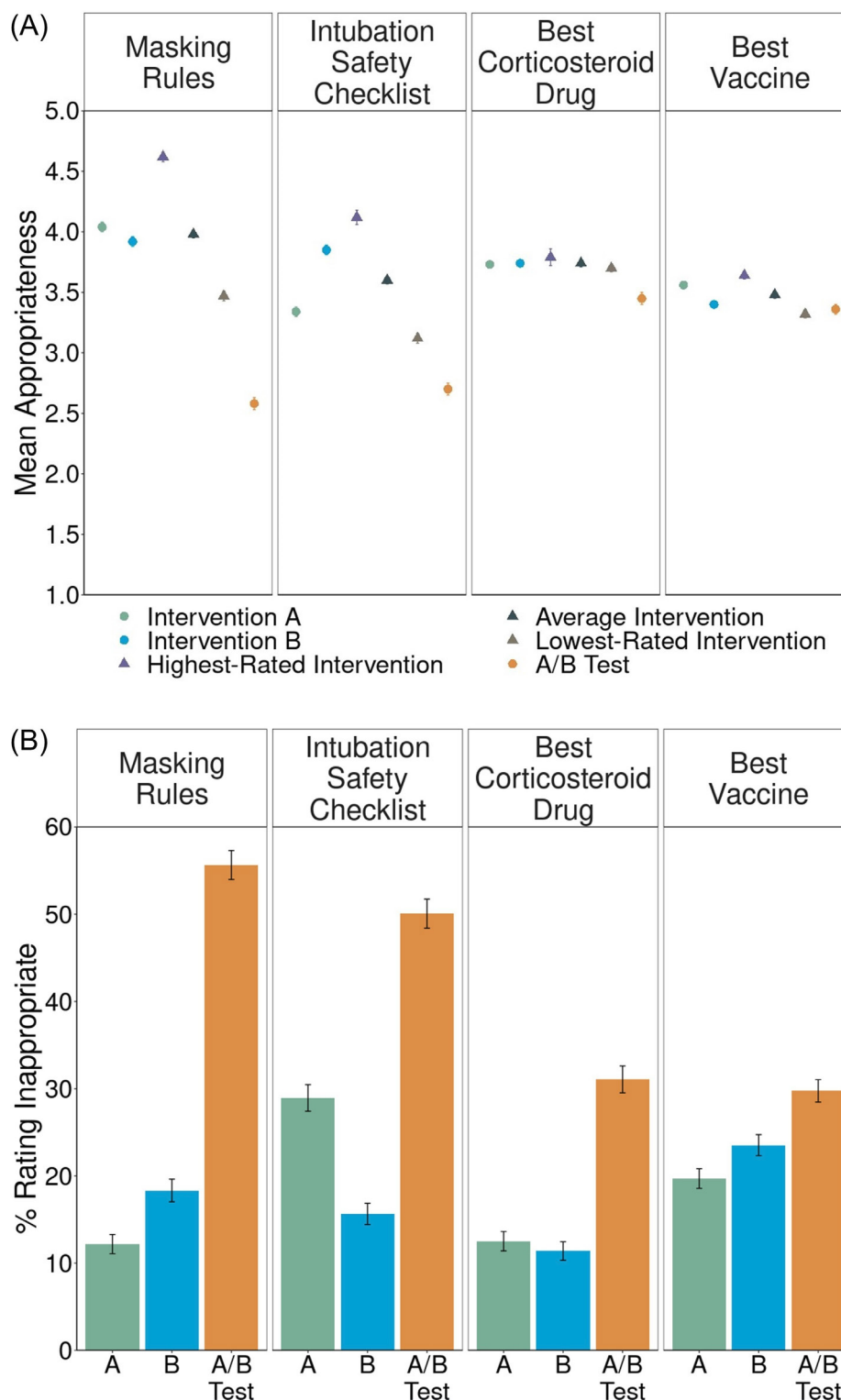


Figure 3 Clinician sentiments about COVID-19 pragmatic randomised controlled trials (pRCTs). (A) Mean appropriateness ratings, on a 1–5 scale, with standard errors, for intervention A, intervention B, the highest-rated intervention, the average intervention, the lowest-rated intervention and the A/B test. Circles represent measures directly collected from participants. Triangles represent averages derived from the direct measures. The distance of the mean appropriateness of the lowest-rated intervention (brown triangle) minus the mean appropriateness of the A/B test (orange circle) represents experiment aversion. The distance of the mean appropriateness of the average intervention (gray triangle) minus the mean appropriateness of the A/B test (orange circle) represents the A/B effect. The distance of the mean appropriateness of the A/B test (orange circle) minus the mean appropriateness of the highest-rated intervention (purple triangle) represents experiment appreciation. (B) Appropriateness ratings transformed into percentages and standard errors of participants objecting (defined as assigning a rating of 1 or 2—‘very inappropriate’ or ‘somewhat inappropriate’—on a 1–5 scale) to implementing intervention A, intervention B and the A/B test.

the illusion of knowledge—here, the belief that either the participant herself or some expert already does or should know the right thing to do and should simply do it—biases people to prefer universal implementation of interventions to pRCTs.^{18 19} One possible solution is to teach patients that clinicians typically have many options for treating a condition, that often no one knows which option is the best, and that a pRCT is the optimal way to figure that out. Similarly, highlighting unwarranted variation in practice during medical training may help reduce clinicians' negative sentiments towards experiments. Rightly or wrongly, both laypeople and clinicians might (a) appropriately recognise that near the start of a pandemic, no one knows which existing drugs, if any, are safe and effective in treating a novel disease, and that new vaccines need to be tested, yet (b) fail to sufficiently appreciate the level of uncertainty around NPIs like masking, proning and social distancing, which can also benefit from rigorous evaluation. This is consistent with the dearth of RCTs (explanatory or pragmatic) of COVID-19 NPIs⁴¹; of the more than 4000 COVID-19 trials registered worldwide as of August 2021, only 41 tested NPIs.⁴² Explaining critical concepts like clinical equipoise or unwarranted variation in medical and NPI practice might diminish experiment aversion.

Limitations

While our lay participant samples were large, diverse and demographically similar to the general US population (see online supplemental table S4), they may not be perfectly representative of other populations. Similarly, Geisinger, the network of hospitals with which the clinicians were affiliated, may not be representative of all hospitals, specifically in their exposure to research and A/B tests such as those described in our vignettes. Geisinger is primarily comprised of teaching hospitals, and has a medical school, but is not associated with a university and, therefore, our results may not generalise either to clinicians who practice at large academic medical centres (eg, Massachusetts General Hospital or Johns Hopkins Hospital) where RCTs are often conducted or, on the other hand, to clinicians who practice at small community hospitals that have little exposure to research. In addition, because the clinician sample was largely made up of individuals with only some research training and experience, these results may not generalise to clinicians who have extensive research training and experience and conduct RCTs (or pRCTs) themselves. Similarly, a large proportion of the clinician sample were nurses and thus the level of experiment aversion observed in these studies may not be representative of the views of physicians and advanced practitioners. Importantly, however, the support of nurses and non-investigator clinical and operational leaders is often needed to conduct a pRCT, and these groups do not always have substantial research experience. Moreover, in both samples, our primary goal was not to estimate the percentage of people in the relevant population who hold negative views of pRCTs, but

rather to ascertain experimentally whether laypeople and clinicians display the patterns of negative sentiments about pRCTs that we have found previously,^{18 19} when confronted with vignettes during, or about, a novel situation (the COVID-19 pandemic). Thus, though the sample may not perfectly represent all healthcare professionals or members of the general public, the results demonstrate the repeated presence of negative sentiments, and a lack of positive sentiments, towards experiments across eight distinct situations among segments of populations whose opinions matter.

Furthermore, because experiment aversion and appreciation are likely sociocultural phenomena, we should expect that the presence or size of the effects we report may differ among societies and over time.⁴³ However, contrary to recent claims,⁴⁴ the similarity in aversion to experiments between laypeople and clinicians suggests that these results generalise across populations that differ in their level of knowledge of RCTs. In addition, our findings here and elsewhere^{18 19} show that experiment aversion occurs in health and non-health scenarios and, within the health domain, in both clinical and public health scenarios, and regarding both pharmaceutical and non-pharmaceutical interventions.

Finally, as noted above, all vignettes discussed in this paper are silent about whether the consent of patients and/or clinicians would be obtained. Previous work that did not directly compare judgments about pRCTs versus treatment implementation suggests that when given the option, laypeople prefer to be asked for consent (eg, for a study comparing the effectiveness of two marketed hypertension drugs, a scenario somewhat related to one of ours^{45 46}). Additionally, other research has found neither experiment aversion nor appreciation (as we define it here and elsewhere³³) after introducing a critical element of voluntariness by asking respondents how likely they would be to 'choose to be treated' at a hospital that is conducting a pRCT.⁴⁴ In separate work, we found that when vignettes explicitly specify that prior consent is obtained, negative sentiment towards pRCTs is reduced—but not eliminated.³³ However, individual consent would undermine the external validity of pRCTs and is anyhow rarely feasible in such settings,^{32 47 48} for example, in tests of policy interventions such as providing safety checklists and promulgating public health rules.

CONCLUSION

Critics rightly note that RCTs have limited external validity when they employ overly selective inclusion/exclusion criteria or are executed in ways that deviate from how interventions would be operationalised in diverse, real-world settings. However, the solution is not to abandon randomised evaluation, but to incorporate it into routine clinical care and healthcare delivery via pRCTs.^{1 48–50} It has been many years since the US Institute of Medicine urged research of many varieties to be embedded in care.⁵¹ More recently, the UK Royal College of Physicians

and National Institute for Health and Care Research issued a joint position statement similarly advocating the integration of research into care.⁵² In addition, the US Food and Drug Administration now promotes pRCTs to support postmarketing monitoring and other regulatory decision-making,^{53,54} a priority also highlighted in the UK Medicines and Healthcare Products Regulatory Agency's 2021–2023 Delivery Plan⁵⁵ and guidance on RCTs.⁵⁶ Pragmatic RCTs have been fielded successfully and informed healthcare practice and policy,^{47,57,58} but they remain far from ubiquitous and they require buy-in to be successful, as shown by the case of a Norwegian school reopening trial during the pandemic that was abandoned due to lack of such support.^{59,60} Broadening the use of pRCTs will require not only redoubling investment in interoperable electronic health records and recalibrating regulators' views of the comparative risks of research versus idiosyncratic practice variation¹ but also anticipating and addressing experiment aversion among patients and healthcare professionals. Better understanding experiment aversion and then discovering strategies to mitigate it will help grow the evidence base necessary for evidence-based decision-making and, ultimately, improved patient outcomes.

Acknowledgements We thank Daniel Rosica and Tamara Gjorgjieva for excellent research assistance.

Contributors PRH, PH, CFC and MNM designed the studies and collected the data. PRH and RLV analysed the data. RLV, RMM, CFC and MNM wrote the first draft of the manuscript. PRH and PH provided critical revisions. RLV and PRH contributed equally to this work. MNM and CFC contributed equally to this work. MNM and CFC are responsible for the overall content as guarantors.

Funding Supported by Office of the Director, National Institutes of Health (NIH) (3P30AG034532-13S1 to MNM and CFC) and funded by the Food and Drug Administration (FDA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the FDA. Also supported by Riksbankens Jubileumsfond grant—Knowledge Resistance: Causes, Consequences, and Cures! to Stockholm University (to CFC), via a subcontract to Geisinger Health System.

Competing interests The authors are or were, at the time the research was conducted, affiliated with the same nonprofit health system (Geisinger) from which the clinician sample was recruited.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval Geisinger's IRB determined that the study surveys were exempted from ethical approval (IRB# 2017-0449), including any requirement of informed consent, under 45 C.F.R. § 46.104(2)(i) (IRB# 2017-0449). Nevertheless, prospective participants were invited to take a survey and told the broad topic, the estimated time it would take and the compensation offered. Those who proceeded were deemed to have tacitly consented. Participants could quit the survey at any time.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement Data are available in a public, open access repository. Participant response data, preregistrations, materials, and analysis code have been deposited at OSF (<https://osf.io/6p5c7/>).

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability

of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Randi L Vogt <http://orcid.org/0000-0003-1709-0471>
 Patrick R Heck <http://orcid.org/0000-0003-0819-3890>
 Rebecca M Mestechkin <http://orcid.org/0009-0002-2976-0364>
 Pedram Heydari <http://orcid.org/0000-0002-9804-1091>
 Christopher F Chabris <http://orcid.org/0000-0002-7379-7378>
 Michelle N Meyer <http://orcid.org/0000-0001-5497-8803>

REFERENCES

- 1 Fanaroff AC, Califf RM, Harrington RA, *et al*. Randomized trials versus commonsense and clinical observation: trials versus commonsense and clinical observation: JACC review topic of the week. *J Am Coll Cardiol* 2020;76:580–9.
- 2 Young SS, Karr A. Deming, data and observational studies. *Signif (Oxf)* 2011;8:116–20.
- 3 New England Journal of Medicine. Introducing NEJM AI. NEJM AI. Available: <https://ai.nejm.org/> [Accessed 28 Feb 2023].
- 4 Grote T. Randomised controlled trials in medical AI: ethical considerations. *J Med Ethics* 2022;48:899–906.
- 5 Loudon K, Treweek S, Sullivan F, *et al*. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;350:h2147.
- 6 Carlo WA, Finer NN, Walsh MC, *et al*. Target ranges of oxygen saturation in extremely preterm infants. *N Engl J Med* 2010;362:1959–69.
- 7 Bilimoria KY, Chung JW, Hedges LV, *et al*. National cluster-randomized trial of duty-hour flexibility in surgical training. *N Engl J Med* 2016;374:713–27.
- 8 Silber JH, Bellini LM, Shea JA, *et al*. Patient safety outcomes under flexible and standard resident duty-hour rules. *N Engl J Med* 2019;380:905–14.
- 9 Rosenbaum L. Leaping without looking—duty hours, autonomy, and the risks of research and practice. *N Engl J Med* 2016;374:701–3.
- 10 Magnus D, Caplan AL. Risk, consent, and SUPPORT. *N Engl J Med* 2013;368:1864–5.
- 11 Rettner R. Premie study triggers debate over informed consent. NBC News; 2013. Available: <https://www.nbcnews.com/id/wbna52439269>
- 12 Carome MA, Wolfe SM. RE: the surfactant, positive pressure, and oxygenation randomized trial (support). Public Citizen; 2013. Available: <https://www.citizen.org/wp-content/uploads/migration/2111.pdf>
- 13 Rice S. Studies on resident work hours —highly unethical, I lack patient consent. Modern Healthcare; 2015. Available: <https://www.modernhealthcare.com/article/20151119/NEWS/151119854/studies-on-resident-work-hours-highly-unethical-lack-patient-consent>
- 14 Bernstein L. Some new doctors are working 30-hour shifts at hospitals around the U.S. Washington Post; 2015. Available: https://www.washingtonpost.com/national/health-science/some-new-doctors-are-working-30-hour-shifts-at-hospitals-around-the-us/2015/10/28/ab7e8948-7b83-11e5-beba-927fd8634498_story.html
- 15 Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci U S A* 2014;111:8788–90.
- 16 Strauss V. Analysis | pearson conducts experiment on thousands of college students without their knowledge. Washington Post; 2018. Available: <https://www.washingtonpost.com/news/answer-sheet/wp/2018/04/23/pearson-conducts-experiment-on-thousands-of-college-students-without-their-knowledge/>
- 17 Hern A. OKCupid: we experiment on users. everyone does. The Guardian; 2014. Available: <https://www.theguardian.com/technology/2014/jul/29/okcupid-experiment-human-beings-dating>
- 18 Meyer MN, Heck PR, Holtzman GS, *et al*. Objecting to experiments that compare two unobjectionable policies or treatments. *Proc Natl Acad Sci U S A* 2019;116:10723–8.
- 19 Heck PR, Chabris CF, Watts DJ, *et al*. Objecting to experiments even while approving of the policies or treatments they compare. *Proc Natl Acad Sci U S A* 2020;117:18948–50.

- 20 Mislavsky R, Dietvorst BJ, Simonsohn U. The minimum mean paradox: a mechanical explanation for apparent experiment aversion. *Proc Natl Acad Sci U S A* 2019;116:23883–4.
- 21 Meyer MN, Heck PR, Holtzman GS, et al. Reply to Mislavsky et al.: sometimes people really are averse to experiments. *Proc Natl Acad Sci U S A* 2019;116:23885–6.
- 22 Dunn A. There are Already 72 Drugs in Human Trials for Coronavirus in the Us. with Hundreds More on the Way, a Top Drug Regulator Warns We Could Run Out of Researchers to Test Them All. *Business Insider*; 2020. Available: <https://www.businessinsider.com/fda-woodcock-overwhelming-amount-of-coronavirus-drugs-in-the-works-2020-4>
- 23 London AJ, Kimmelman J. Against pandemic research exceptionalism. *Science* 2020;368:476–7.
- 24 Dominus S. The covid drug wars that pitted doctor vs. doctor. The New York Times; 2020. Available: <https://www.nytimes.com/2020/08/05/magazine/covid-drug-wars-doctors.html>
- 25 Qualtrics XM: The Leading Experience Management Software, Available: <https://www.qualtrics.com/> [Accessed 24 Apr 2024].
- 26 CloudResearch, Available: <https://www.cloudresearch.com/> [Accessed 24 Apr 2024].
- 27 Litman L, Robinson J, Abberbock T. TurkPrime.com: a versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav Res* 2017;49:433–42.
- 28 Amazon Mechanical Turk, Available: <https://www.mturk.com/> [Accessed 24 Apr 2024].
- 29 Germine L, Nakayama K, Duchaine BC, et al. Is the Web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychon Bull Rev* 2012;19:847–57.
- 30 Simons DJ, Chabris CF. Common (mis)beliefs about memory: a replication and comparison of telephone and mechanical turk survey methods. *PLOS ONE* 2012;7:e51876.
- 31 Créquit P, Mansouri G, Benchoufi M, et al. Mapping of crowdsourcing in health: systematic review. *J Med Internet Res* 2018;20:e187.
- 32 Asch DA, Ziolek TA, Mehta SJ. Misdirections in informed consent — impediments to health care innovation. *N Engl J Med* 2017;377:1412–4.
- 33 Vogt RL, Mestechkin RM, Chabris CF, et al. Objecting to consensual experiments even while approving of nonconsensual imposition of the policies they contain *PsyArXiv* [Preprint] 2023.
- 34 Greene JD, Sommerville RB, Nystrom LE, et al. An fMRI investigation of emotional engagement in moral judgment. *Science* 2001;293:2105–8.
- 35 Dunlap WP, Cortina JM, Vaslow JB, et al. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods* 1996;1:170–7.
- 36 Westfall J. Effect Size | Cookie Scientist, 2016. Available: <http://jakewestfall.org/blog/index.php/category/effect-size/> [Accessed 30 Mar 2023].
- 37 Vogt RL, Heck PR, Mestechkin RM, et al. Data from: aversion to pragmatic randomized controlled trials: three survey experiments with clinicians and laypeople in the united states. OSF Repository; 2024. Available: <https://osf.io/6p5c7/>
- 38 Angus DC. Optimizing the trade-off between learning and doing in a pandemic. *JAMA* 2020;323:1895–6.
- 39 Goodman JL, Borio L. Finding effective treatments for COVID-19: scientific integrity and public confidence in a time of crisis. *JAMA* 2020;323:1899–900.
- 40 Manzi J. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books, 2012.
- 41 McCartney M. We need better evidence on non-drug interventions for covid-19. *BMJ* 2020;370:m3473.
- 42 Hirt J, Janiaud P, Hemkens LG. Randomized trials on non-pharmaceutical interventions for COVID-19: a scoping review. *BMJ Evid Based Med* 2022;27:334–44.
- 43 Bas B, Vosgerau J, Ciulli R. No evidence that experiment aversion is not a robust empirical phenomenon. *Proc Natl Acad Sci U S A* 2023;120:e2317514120.
- 44 Mazar N, Elbaek CT, Mitkidis P. Experiment aversion does not appear to generalize. *Proc Natl Acad Sci USA* 2023;120:e2217551120.
- 45 Cho MK, Magnus D, Constantine M, et al. Attitudes toward risk and informed consent for research on medical practices. *Ann Intern Med* 2015;162:690–6.
- 46 Nayak RK, Wendler D, Miller FG, et al. Pragmatic randomized trials without standard informed consent?: a national survey. *Ann Intern Med* 2015;163:356–64.
- 47 Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med* 2019;381:1175–9.
- 48 Wieseler B, Neyt M, Kaiser T, et al. Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy? *BMJ* 2023;380:e073100.
- 49 Simon GE, Platt R, Hernandez AF. Evidence from pragmatic trials during routine care - slouching toward a learning health system. *N Engl J Med* 2020;382:1488–91.
- 50 Morales DR, Arlett P. RCTs and real world evidence are complementary, not alternatives. *BMJ* 2023;381:736.
- 51 Olsen L, Aisner D, McGinnis JM. *IOM Roundtable on Evidence-Based Medicine, The Learning Healthcare System: Workshop Summary*. Washington, DC: National Academies Press, 2007.
- 52 RCP nihr position statement: making research everybody's business. RCP London; 2022. Available: <https://www.rcplondon.ac.uk/projects/outputs/rcp-nihr-position-statement-making-research-everybody-s-business>
- 53 Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence - what is it and what can it tell us? *N Engl J Med* 2016;375:2293–7.
- 54 Office of the Commissioner. Real-world evidence. FDA; 2023. Available: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>
- 55 The medicines and healthcare products regulatory agency delivery plan 2021-2023. GOVUK; 2022. Available: <https://www.gov.uk/government/publications/the-medicines-and-healthcare-products-regulatory-agency-delivery-plan-2021-2023> [Accessed 26 Oct 2023].
- 56 MHRA guideline on randomised controlled trials using real-world data to support regulatory decisions. GOVUK; 2023. Available: <https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions/mhra-guideline-on-randomised-controlled-trials-using-real-world-data-to-support-regulatory-decisions> [Accessed 22 Jan 2024].
- 57 Finkelstein A, Zhou A, Taubman S, et al. Health care hotspotting - a randomized, controlled trial. *N Engl J Med* 2020;382:152–62.
- 58 Weinfurt KP, Hernandez AF, Coronado GD, et al. Pragmatic clinical trials embedded in healthcare systems: generalizable lessons from the NIH Collaboratory. *BMC Med Res Methodol* 2017;17:144.
- 59 Fretheim A. ISRCTN44152751: school opening in norway during the covid-19 pandemic. ISRCTN registry. 2020 Available: <https://www.isrctn.com/ISRCTN44152751>
- 60 Fretheim A, Flatø M, Steens A, et al. COVID-19: we need randomised trials of school closures. *J Epidemiol Community Health* 2020;74:1078–9.