**The Fault lies in our Stars not in Ourselves:
Randomized Controlled Trials and Clinical Knowledge?**

**David Healy – August 2020**

**Comments from Lisa Bero (indented)**

**Executive Summary**

As a matter of the historical record, randomized controlled trials (RCTs) stumbled into being without a coherent clinical underpinning.

I would argue that RCTs themselves are not incoherent… they were never expected to have external validity. RCTs as a method have a lot of value.. not sure if you have read the book below. I would start with your point #4 below – the misuse of RCTs by industry. This to me focuses the problem and would make your average methodologist take less offense. I have written about the misuse of RCTs and methods base on them (including systematic reviews) in environmental health. So, it is not the method that is the problem, it’s the application.

[Randomistas by Andrew Leigh - Black Inc Books](https://www.blackincbooks.com.au/books/randomistas)

You base your arguments on examples, and I can think of good examples of RCTs as well as bad examples… so, again, this is why I would not disparage the method, but it’s use.

RCTs have proliferated for reasons of bureaucratic convenience rather than epistemological coherence.

This could be.. would use FDA and regulatory agencies as an example. Esp when it comes to their obligation to evaluate harm

Elements of RCTs, such as randomization, confidence intervals, and primary endpoints, can help in treatment evaluation but their indiscriminate combination can cause problems.

I didn’t really understand this section… not sure what you mean by indiscriminate combination

Pharmaceutical company use of RCTs for regulatory and marketing purposes gives rise to another set of problems distinct from the ones outlined here.

I would use the above as your starting point… what are the problems?

The rhetoric pitching RCTs as offering gold-standard evidence misleads as regards both treatment hazards and benefits

I would agree with above, there was an interesting article in NEJM suggesting the term gold standard did not come into use until the 80s. I have written about, in public health, having the best evidence to suit your question – thus, no methods is a gold standard. It will depend on the question.

Treatment induced morbidity and mortality is rising, possibly driven by the use of RCTs as our primary treatment evaluation tool.

I think you will have trouble establishing causality!

**In the Beginning**

In 1947, a trial of streptomycin introduced RCTs to medicine. From then, through to their incorporation into the 1962 amendments to the Food, Drugs and Cosmetics Act, occasioned by the thalidomide tragedy, doubts were expressed as to whether RCTs had a solid epistemological basis in clinical reality. Since then there have been ongoing disputes about the best statistical approach to take to RCT data – whether confidence intervals are preferable to significance testing, for instance. There have also been efforts to account for a heterogeneity of treatment effects (HTE) within the wider Evidence Based Medicine (EBM) movement, which touch on the issues raised here (Kravitz et al 2004). But few doubts have been expressed about the fundamental validity of RCTs.

Repeated characterizations of this form of trial as offering gold-standard evidence likely leave most clinicians thinking RCTs have a solid epistemological foundation, even as they recognize difficulties in translating from population or average effects to individual patients. If clinicians have no apparent qualms, few others likely feel compelled to dig deeper. Even those who recognize problems accept a characterization of RCTs as generating evidence that is generalizable and knowledge that lies within confidence limits in contrast to the views of clinicians and case reports.

**Hill and Fisher**

A Medical Research Council (MRC) trial of streptomycin in 1947 tested the feasibility of randomization as a further control of the subtle biases involved in evaluating a medicine. This RCT demonstrated that randomization was feasible but the information it produced was less clinically relevant than a prior trial of streptomycin that controlled for confounders in the then standard way and depended on clinical judgement (Healy 2012; Healy 2020). The standard trial noted important therapeutic issues the RCT missed - streptomycin could cause deafness and tolerance developed rapidly.

Tony Hill, the MRC trial lead, did not consider the epistemological validity of RCTs. He had taken the idea of randomization from a horticultural thought experiment Ronald Fisher outlined two decades previously in which Fisher proposed that randomization could control for unknown confounders. Hill’s idea was that it might control for the difficult to detect ways in which clinicians steer patients likely to respond well into an active treatment arm. There was no consideration as to whether randomization would control for the unknown of doctors not knowing what they were doing (Healy 2020).

RCTs brought statistical significance tests in their wake because Fisher tied randomization to significance testing. The only things that can interfere with expert judgement not being correct every time, Fisher argued, are unknown confounders and chance. Significance testing could control for chance and randomization could control unknown confounders. Fisher’s model had an anchor in the real world – an expert who knew what they were doing and whose judgements were invariably correct – such as offering a view that wearing a parachute if you jump from a plane at 5,000 feet will save your life. For Fisher, experiments were a way of demonstrating that we knew what we were doing rather than a leap into the unknown - they should get the same result every time.

There are situations in medicine, where doctors know what they are doing. The more they know, the more they approach Fisher’s expert, but no one runs randomized trials in situations where we are likely to get the predicted result every time.

In the case of breast cancer, on the basis of advances in physiology, it was hoped that giving Herceptin to Her 2+ receptor breast cancers might produce better responses than cisplatin, a more indiscriminate toxin, which nevertheless extends longevity compared to placebo. Trials confirm this but also reveal that even using Herceptin in Her 2+ breast cancers, we do not get the same result every time – there is a lot we don’t know.

Our lack of knowledge is even more marked in trials testing stents compared to other cardiac procedures, where doing what seems obvious does not produce the expected results. The issue is not whether stents work or not but whether we know what we are doing, which we mostly don’t. While recent stent RCTs are a good demonstration of the power of RCTs to stall a therapeutic bandwagon, the view that clearing blocked arteries might not produce a good outcome had been accepted clinical wisdom in vascular leg surgery for decades and for stents in some quarters before any RCT had been run.

**Confidence Intervals**

Taking issue with Fisher’s real-world anchor - an expert knowing what they were doing - Jerzy Neyman and Egon Pearson borrowed from Carl Friedrich Gauss’ use of confidence intervals to manage the error in astronomical measurements of stars. Gauss’ ideas were picked up by Pierre-Simon Laplace and their combined input (1809-1827) to the central limit theorem, least-squares optimization and efficient linear algebra provided celebrated benefits for the physical sciences, engineering, astronomy, and geodesy. When applied to the problem of imprecise measuring instruments and invariant entities like stars, confidence intervals have an anchor in the real world, helping us to decide if our varying measures reflect the presence of one or two stars.

Confidence intervals unquestionably work but are not necessarily valid in clinical trials. Taking successive measurements of a pulse in an individual is similar to determining the precise location of a star – the tighter the confidence interval bounding our measurements the more apparent we can do things reliably. Using confidence intervals to describe data also make it less likely we will reject a drug like Herceptin just because we don’t get the same result every time.

Confidence intervals could be used in a manner consistent with their use in astronomy to distinguish between a repeated set of pulse measurements before and after administration of a drug - to one individual. The current use of confidence intervals in RCTs seems predicated on the idea that a cohort of patients in standard parallel group trials can be regarded as a single object like a galaxy. But pulses can increase in response to a drug in one individual and decrease in a second in response to the same drug – this is not measurement error.

Claiming the true effect of the drug likely lies near some mean of the effects in a group of individuals, potentially giving us a best estimate of no effect, is wrong. When a mechanism to decide whether there are one or two stars present turns up the answer there are none, something has gone wrong. Distinguishing between Average Treatment Effects (ATE) and HTE recognizes but doesn’t solve this problem. A series of trial designs have been proposed to mitigate the problem, with some recognition that the degree to which the problem can be mitigated depends on how often individuals respond in opposite ways – if commonly the notion of average treatment effects falls apart. In the case of stars, we knew enough about what we were doing to make reasonable inferences from varying measurements. We need to know as much to make comparable inferences when giving medicines.

**Primary Endpoints**

There is a deeper problem than just how commonly people react differently. Fisher’s significance testing, and Gauss’ confidence intervals, require a focus on a primary endpoint – are there one or two stars or more or less ears of corn? In medical RCTs, as currently undertaken, a focus on a primary endpoint is viewed as key to ensuring that only chance or measurement error will get in way of the correct result. This rather than the use of confidence intervals or significance testing per se limits the value of RCTs in medicine.

A horticultural expert focused on whether a fertilizer improved corn yield would likely have no more accurate a view of its effects on worms in the ground or insects in the air than a non-horticulturalist - in respect of whose views significance testing by Fisher’s definition would not be appropriate.

Similarly, Gauss’ confidence intervals applied to an astronomer’s measurements of the location of a star are of little use when it comes to pinpointing the trajectories of satellites crossing the path of the observations.

There is something of an assumption that the primary endpoint in an RCT is the commonest effect of a drug. Treatment heterogeneity leading to wider confidence intervals than are ideal can be accommodated against this background as can missing other effects assumed to be rare or not appearing within the duration of the trial. But RCTs are often not trained on the key or commonest thing a drug does. The commonest effect of an SSRI is genital anesthesia, which appears almost universally and within 30 minutes of taking a first pill. It should not be possible to miss it, but this effect has been missed in all RCTs of these drugs for nervous conditions because of an RCT required focus on a primary endpoint. A focus on a primary endpoint is essentially an act of hypnosis that likely routinely leads to common treatment effects being missed entirely.

RCTs can help evaluate one effect of a drug but are a poor method to evaluate a drug’s overall effects. The precision focus on a primary endpoint means we might end up being able to say a lot about a pimple but very little about the person on whose back it is. Casting RCTs as offering gold standard evidence about a drug creates an ignorance about the ignorance they generate.

A negative result in an RCT may temper the enthusiasm for an apparently obvious course of action such as stenting coronary arteries but the stent trials do not provide a basis for dismissing a link between an apparently good outcome in an individual case after stent insertion. The trials demonstrate we do not know what we are doing in general; there may be a subset of patients in which a stent is the correct answer. In an ideal world, the greatest utility of treatment RCTs would not lie in their giving us the right answer but that a negative result should force believers in a treatment to delineate a patient group who clearly benefit.

RCT evidence should never trump an evident effect that appears after treatment. If a person becomes suicidal after taking an antidepressant, the issue of what is happening in that case is a matter of assessing the effects of their condition, circumstances, prior exposure to similar drugs, dose changes on the medication and whether there are other evident effects of treatment consistent with a link between suicidality and treatment. Unless RCTs have been designed specifically to look at the effects of treatment on a possible emergence of suicidality (and there have been none), RCT evidence is irrelevant.

**Confounding by Ignorance**

In the case of RCTs, inconvenient results often come with a rider that confounding by indication may make their interpretation difficult. This translates into a caution against assuming a treatment is causing a problem. Many clinicians and the wider public will not be clear why randomization does not take care of confounding by all unknown unknowns – such as confounding by indication. Re-designating this as confounding by ignorance may better convey why not.

Consider two scenarios involving the antidepressants, imipramine and paroxetine. Imipramine was discovered in 1957 and launched in 1958 without any RCT input. Among other actions, it is a serotonin reuptake inhibitor. In later RCTs it (and other older antidepressants all discovered and marketed without RCTs), “beat” SSRI antidepressants in trials involving patients with melancholia (severe depression). Melancholic patients are 80 times more likely to commit suicide than mildly depressed patients.

By 1959, clinicians noted that imipramine, wonderful though it was, could cause agitation and suicidality in some patients that cleared when the drug was stopped and reappeared when restarted – Challenge-Dechallenge-Rechallenge (CDR) proof it causes suicide.

In an RCT of imipramine in melancholia, a drug that can cause suicide, imipramine seems likely to protect against suicide on average by reducing the risk from melancholia to a greater extent than placebo. In contrast, in the RCTs that brought SSRIs to the market, these drugs doubled the rate of suicidal acts. This was because SSRIs are weaker than imipramine and had to be tested in people with mild depression, at little risk of suicide. The low placebo suicidal act rate revealed the risk from the SSRI – as it does for imipramine put into trials of mild depression. RCTs can, in other words, mislead as regards cause and effect – potentially getting results all the way along a spectrum from “causes”, to possible risk, likely protective and “cannot cause”.

In any trial where both condition and treatment cause superficially similar problems, as when antidepressants and depression cause suicidality or bisphosphonates and osteoporosis both lead to fractures, a dependence on RCT data rather than clinical judgement risks misleading. This is likely the case for a majority of RCTs in clinical conditions, which are Treatment Trials rather than Drug Trials.

Drug Trials are done on healthy volunteers, and in these, which ordinarily do not have a primary endpoint, treatment effects stand out more clearly. SSRI Drug Trials demonstrated sexual effects were common, often debilitating, and might endure after treatment stopped, that agitation up to suicidality was common and that dependence commonly occurred after exposures of two weeks. The correct choice of primary endpoint in subsequent Treatment Trials could eliminate these effects – provided the non-confidential Drug Trial data remained unpublished.

Paroxetine was later put into Treatment Trials of patients with Major Depressive Disorder (MDD) and patients with Intermittent Brief Depressive Disorders (IBDD). IBDD patients (borderline personality disorder) are repeated self-harmers. Their depressive features, however, mean that IBDD patients can readily meet criteria for MDD.

In April 2006, GlaxoSmithKline (GSK) released RCT data showing MDD patients had a worrying increase in suicidal event risk on paroxetine (Table). The data from IBDD RCTs in the GSK release were better. We can add 16 suicidal events to the paroxetine IBDD column and still get an apparently protective rather than problematic result for paroxetine when MDD and IBDD data are added together.

**Table: Suicidal Events in MDD & IBDD Trials**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Paroxetine** | **Placebo** | **Relative Risk** |
| MDD Trials Acts/PatientsIBDD Trials Acts/PatientsCombined Acts/Patients | **11/2943****32/147****43/3090** | **0/1671****35/151****35/1822** | **Inf (1.3, inf)****0.9****0.7** |

This effect has been noted as a hazard of meta-analyses and is termed Simpson’s paradox but it has to apply to some extent in every trial that recruits patients who have a superficially similar but in fact heterogenous condition such as depression, pain, breast cancer, Parkinson’s disease or diabetes – almost all medical disorders. The issue here is our ignorance about the heterogeneity of conditions rather than an additional HTE problem.

Every time there is a mixture of more than one patient group in a trial, randomization will ensure some patients hide some treatment effects – good and bad. Trials of standard treatments for back pain, for instance, will mask the beneficial treatment effects of an antibiotic on back pains linked to infections (up to 10% of back pains). It will often be possible to design a Treatment Trial that uses an effect a drug causes to hide that same effect.

This confounder applies to placebos. The assumption is that in Treatment Trials placebos simply control for natural variation. But placebos can have potent treatment effects, making them another treatment like an antibiotic in the backpain scenario above. We do not know enough about placebo responses to know whether in the context of randomization they control or confound the data.

Every medicine that gets on the market, by definition, beats placebo (often inconsistently). As a result, it has become unethical to use placebos in clinical practice, when for some patients it is possible to do open heart surgery on placebo and for those for whom it works more generally a placebo may be a better idea than therapeutic poisoning.

Another problem is that a quantitative approach to data generated by algorithm rather than an approach based on judgement increases the risk that minor events in a placebo arm will be offset against significant events in an “active” treatment arm offering an opportunity to claim that nothing specific has happened, when it has.

**“Side” Effects**

The suicidality, sexual dysfunction, agitation, or insomnia caused by antidepressant in clinical trials are commonly folded into a primary endpoint, the Hamilton Depression Rating Scale (HDRS), which includes questions on suicidality, sexuality, sleep and agitation. These changes render confidence intervals around scores on these items meaningless, compromise the use of the scale more generally, and risk hiding a benefit.

If registered on adverse event forms, treatment emergent suicidality or sexual dysfunction should almost de facto be causally linked to treatment. But investigators are forced to tick boxes as to the likelihood of a link, including possibly unrelated, toward which the ethos of RCTs, which aims at replacing clinical judgement with decisions based on an analysis of datapoints rather than an interrogation of people, steers them.

The question of whether the suicidality the patient in front of me is experiencing comes from their illness or their treatment is not a matter of deciding if there are 1 or 2 stars. In this case, we already know there are two stars and a lot about them. Instruments specifically designed with the characteristics of each star in mind may facilitate the distinction, but ultimately it’s a case of pattern recognition and a judgement call – with one call leading to an increase in the dose of treatment and the other leading to a reduction. The high stakes may make the option of falling back on an algorithm appealing – but it is not good science or good medicine.

In 1959, clinicians could readily distinguish between a treatment emergent suicidality on imipramine and the suicidality caused by melancholia. In 1961, Frank Ayd, the discoverer of amitriptyline could distinguish the sexual dysfunction it causes from the sexual dysfunction melancholia causes. Through to 1991, clinical knowledge of drug effects derived primarily from clinical experience, embodied in case reports and published in clinical journals. The steady rise of mechanical evaluations, however, allied to a sequestration of trial data, has relegated clinical evaluations that drug X causes effect Y, even when buttressed by evidence of CDR, to the status of anecdotes. From 1991, journals stopped taking anecdotes about “side” effects that almost by definition are rare compared to the treatment effect; the bad luck of a few patients is not viewed as deserving to be weighed in the balance against science.

As a result, where in the 1960s the harms of treatments took at most a few years to establish after a drug came on the market, by 1990 a set of processes were in place that mean it can now takes decades for the harms of treatments to be established, as for instance in the case of impulse control disorders on dopamine agonists or persistent sexual dysfunction on isotretinoin, antibiotics, analgesics, finasteride and other drugs.

This growing delay underpins a perception that pharmacovigilance is in crisis. Proposed solutions mention the need for systems to detect the rare effects of treatments not found in RCTs. There is a turn to a mining of electronic medical records or other observational approaches. New signal detection methods and investigative approaches are always welcome, but these are not the answer to the problems we now face, which lie not in a failure to detect rare effects but in a systematic failure to acknowledge common effects. Drug Trials are a better way than any new signal detection method to detect these common effects.

The ability of RCTs to focus on one effect suits trials testing for an effect of possible commercial interest. But this focus does not suit an evaluation of treatments, the intention of which is to poison or mutilate in the hope of producing an overall benefit. Studies run on a primary endpoint chosen for commercial reasons cannot be expected to produce the kind of information that might inform therapeutic poisoning. Can we expect data-handling methods developed for fertilizers and stars to encompass the complexity of therapeutic poisoning?

Commercial trials produce regulatory pronouncements that a treatment’s Risk-Benefit ratio is favorable. A favorable Risk-Benefit ratio here implies there is a balance in which all benefits and risks can be weighed. There is no balance, and there is no weighing. One statistically significant effect is taken to count for more than all other effects, even serious effects that occur more frequently and can include death, but which by design are not significant.

**Objectivity**

Concerns about what is often termed the population effects of RCTs and the mismatch between these and the apparently “anomalous” responses of individual patients has been framed in terms of HTE and recognized in EBM as needing an incorporation of RCT evidence into the judgement of clinicians and the values and preferences of patients.

The argument here is that RCTs, as they have evolved RCTs, are essentially algorithmic or operational. DSM criteria in mental health, and the metrics for blood pressure, peak flow rates and bone densities are similarly operational. The creators of the DSM criteria claimed that “of course just ostensibly meeting criteria for an illness doesn’t mean the person has the illness” – clinical judgements are needed to establish what is really going on in the case of criteria, just as they are in the case blood pressure, peak flow or bone density metrics. In practice, however, operational exercises like RCTs, DSM criteria and many medical metrics nudge us toward a suspension of judgement and put a third party like the pharmaceutical industry in a strong position to contest any introduction of judgement by a doctor or patient on the basis that the figures are demonstrably more objective than any clinician or patient judgement can be

Science traditionally generates data and challenges us to interpret them. New techniques (like a new drug) can throw up new observations, but while new data can challenge prior judgements, the mission of science has not been to replace judgement by technique.

Individual judgement of course is suspect. This argument does not advocate replacing collective evaluation by a reliance on individuals or doctors; the argument is for collective evaluation rather than its replacement by algorithmic processes. Collective evaluation has a clear footing in the real world, as the streptomycin trial demonstrated; the idea that RCTs have as clear a footing has been assumed rather than established.

Arguments stressing the need for RCTs rather than clinical judgement point to a small series of treatments, such as internal mammary ligation, that RCTs demonstrated did not work, with the implication that clinical judgement can get things wrong. The RCT in this case was only run because the dominant clinical judgement was that these treatments didn’t work – an article in the Reader’s Digest notwithstanding.

These arguments fail to note that most of the current treatment classes we have were introduced in the 1950s without RCTs. That the treatments introduced then from anti-hypertensives and hypoglycemics to psychotropic drugs are more effective than treatments introduced since. That RCTs facilitate the introduction of treatments with lesser effects.

The response to the point about weaker treatment effects is the same as the response to anomalous effects – namely that RCTs give population or average treatment effects. This implies a valid population with individual outliers, when in fact there is no knowing how ***any*** individual will respond to an antidepressant for instance. Fisher expected us to get the same result in every individual case, and within limits confidence intervals offer the same guarantee. Neither Fisher nor Gauss would recognize a problem in translating from a “population” to an individual level. Because of diagnostic imprecision, individual heterogeneity, treatment effect heterogeneity, and other factors we don’t have valid populations. To adapt Shakespeare, the fault here lies in our stars not in ourselves.

Clinical practice is essentially a judicial rather than an algorithmic exercise. The view offered here is that our best evidence as to what happens or is likely to happen on treatment lies in the ability to examine and cross-examine the person or persons (interrogate the data) given that treatment. Every day of the week doctors and patients continue or stop treatments based on judgements as to whether the treatment is working or not. These judgements have to be mostly correct or else medicine would not work.

But what holds true at the individual level must be true at the population level also. The evaluation of a treatment must be judicial rather than algorithmic.

An endorsement of clinical judgement does not suit health service managers or the pharmaceutical industry, for whom the supposed generalizability of RCT knowledge and confidence intervals that can be offered for such knowledge are legally appealing.

Behind these problems lies the undefined concept of clinical data. This paper takes the position that data means the people entered into a study - the people who lie behind any table of figures or behind the outcome of any analytic process applied to those figures. At present case reports with names attached are the only form of controlled clinical investigation that offers access to the data, the possibility to interrogate the data and, accordingly, an opportunity to ground any conclusions in the real world.

**The Place for Randomized Controlled Trials**

RCTs are an important tool to evaluate certain drug effects rather than a means of generating gold-standard knowledge about drugs or treatments.

They came into widespread use after 1962 on the back of claims by Louis Lasagna and Walter Modell they might solve a regulatory problem – not because of their epistemological validity or proven superiority to other methods of evaluation (Healy 2012; Healy 2020).

In 1965, Tony Hill offered the view that RCTs have a place in the study of ***therapeutic efficacy,*** but they are only one way of doing so and any belief to the contrary was mad (Hill 1965).

The subsequent history of company trials demonstrates that the knowledge derived from RCTs is far from generalizable. In 1962, when two positive placebo-controlled trials were put in place as the criterion of entry to the market, it was assumed any subsequent trial would almost always replicate the result of a first positive trial. This is clearly not the case.

Lasagna, the leading advocate for RCTs up to 1962, also changed his view, adding in respect of adverse events that while the common view was that spontaneous reporting was unsophisticated and not scientifically rigorous, this was only the case in the dictionary sense of sophisticated meaning “adulterated” and spontaneous reporting was in fact more worldly-wise, knowing, subtle and intellectually appealing than RCTs (Lasagna 1983).

Randomization, placebo controls, confidence intervals and primary endpoints all have a place in the evaluation of treatments. Confidence intervals are clearly appropriate in instances where measurement error is likely to play a part.

Randomization is an extra control on clinical bias. There is a place for it, unhooked from primary endpoints and statistical significance, as happens in large pragmatic trials – but here the word pragmatic concedes our limited understanding of what we are doing.

There is a place for it in surgical studies of a time-limited intervention as opposed to chronic therapeutic poisoning, as well as in studies to evaluate programs, and treatment studies that have a hard endpoint – like all-cause mortality, but even here we risk being misled by findings of no change in mortality into missing a switch from cardiac events to cancers when many patients might preferred to die by a heart attack (Mangin et al 2007).

RCTs also have a merit as a gateway to the market; randomization means that trials require less patients and can be run quickly. A positive result in commercial trials may indicate a compound has an “effect”. Trials aimed at establishing effectiveness, in contrast, require hard outcomes and time. This is not a realistic gateway to the market. Demonstration of an effect, as with SSRIs for depression, means it is not correct to say this drug does nothing and on this basis entry to the market could be permitted.

By implication, the launch of a drug licensed on these terms would be the point when more comprehensive clinical evaluations should start, aimed at generating a clinical consensus as to the place of the drug in practice. As a general tool to evaluate the effects of a drug, RCTs should take second place to a group of experienced clinicians whose observations are not constrained by checklists and an investigation tailored to one effect.

After 1962, RCTs became the standard through which industry would make gold, As they proliferated, and became increasingly mechanical, the mantra that they provide gold standard medical evidence took hold. The ignorance of ignorance in claims that the only valid information on medicines comes from RCTs compounds a series of other factors that make RCTs a gold standard way to hide adverse events and encourage over-use of treatments.

It is appropriate to use RCTs to raise the bar to those who would make money from one effect of a drug given to people at their most vulnerable, but seasoned clinicians, allied to increasingly health-literate patients, are better placed than RCTs to determine cause in the case of the 99 other effects every drug has, especially effects such as sexual or suicidal effects of antidepressants, for instance, that need input from patients able to distinguish condition effects from superficially similar treatment effects.

Drug interventions (therapeutic poisoning) invariably harm; the hope is that some good can also be brought from their use. Evaluations by RCT harm (generate ignorance), but if used judiciously some good can be brought out of the ignorance they necessarily generate.

It is less likely that good will be brought out of ignorance if we rely solely on a data handling formula. Analytic methods can describe data but whether good comes from their use requires the kind of judgement calls that statistical approaches ordinarily make a virtue of side-lining. A recent study looking at 29 ways to analyze a dataset, generated from referees giving red cards to dark and light-skinned soccer players, demonstrated that different techniques can lead to a wide variation in results with none able to guarantee what is happening in the real world (Silberzahn, Uhlmann, Martin et al. 2018).

**Commercial RCTs**

Pharmaceutical companies run “RCTs” for regulatory and marketing purposes, and this may generate a belief that any problems with RCTs stem from a link to commerce.

The difficulty in recognizing adverse effects has for instance been compounded by company sequestration of clinical trial data and a ghost writing of the clinical literature that hypes the benefits and hides the harms of treatments, compounded by a regulatory willingness to avoid deterring patients from treatment benefits by placing warnings on drugs.

Clinical practice is also compromised by licensing indications and by guidelines. There are no drugs licensed to treat adverse effects. When a person becomes suicidal on an SSRI, there is no treatment licensed to treat this toxicity. Many clinicians wanting to help feel compelled to diagnose depression rather than toxicity but a depression diagnosis inevitably leads to a further treatment with an antidepressant rather than something more appropriate like a benzodiazepine, a beta-blocker or red wine.

Guidelines based on a ghostwritten literature without any access to trial data list treatment benefits, but none tackle treatment harms, and this contributes to their disappearance.

The incorporation of RCTs into a bureaucratic regulatory apparatus has introduced elements such as surrogate markers, which mean that in real life treatments may not show effectiveness consistent with RCT demonstrations of a treatment effect. Trials showing antidepressants work, for instance, bizarrely have an excess of deaths and both suicidal and homicidal events in their treatment arms compared to placebo.

Commercial trials have given rise to the idea of an abstract Risk-Benefit ratio which along with treatment effect sizes, the Number Needed to Treat (NNT) or to Harm (NNH) are not based in clinical reality.

Possible answers to these problems lie with medical journals who should insist on the data from Drug Trials being published perhaps in conjunction with Treatment Trials and should not publish Treatment Trials without full access to their data. Our hierarchies of evidence should come clean on whether they regard a ghostwritten article without access to clinical trial data as better than or inferior to a Case Report that embodies dose responsiveness and CDR elements for instance. And those deploying an analytic process should clarify how the resulting figures translate into the real world, rather than assuming they do.

But these problems avail of and aggravate a pre-existing epistemological gap rather than constitute the gap.

**Coda**

Evaluating the effects of drugs is among the most important exercises we undertake. When drugs work, they can like parachutes save lives. Given the importance of the task, the notion of a hierarchy of evidence at the top of which are mechanisms that do the deciding for us has a potent allure.

Relegating judgement to the bottom of the evidence hierarchy in medicine brings out our discomfort with judgement – both clinical and pharmaceutical. The proliferation of Intention to Treat and related analyses hinge on the same discomfort. Succumbing to an operational solution, however, is at least as dangerous as depending on judgement.

In the case of airplanes, adding parachutes and other interventions that are effective (rather than just have an effect) enhances safety, although recent Boeing plane crashes point to the perils of too great a reliance on automatic decision tools.

RCTs have led many to view drug treatments as comparable in effectiveness to parachutes. As a result, by the age of 50, close to 50% of us are now on five or more drugs. For the past five years, our life expectancies have been falling and admissions to hospital for treatment-induced morbidity are rising, an outcome that contrasts with the added safety of having parachutes and other gadgets in planes (Healy 2020). Current data indicates that stripping out the number of gadgets from 10 or more to 5 or less reduces hospitalization, increases life expectancy and improves quality of life (Garfinkel and Mangin 2010).

While simply combining five pluripotent drug gadgets almost certainly brings risks of interactions that airplane gadgets don’t bring, if current RCTs of medicines essentially produce evidence that it is not correct to say this drug has no possible benefit, it makes sense to think that being on five or more drugs, of which all we can say in the case of each of them is that it is not right to say they have no possible benefit, might be a contributing factor to increasing levels of mortality and morbidity.

Recent data on life expectancies and treatment linked morbidities call for an evaluation of the role of RCTs in the evaluation of drug treatments (Healy 2020).

**References:**

Garfinkel D, Mangin D. Feasibility study of a systematic approach for discontinuation of multiple medications in older adults. Arch Intern Med 2010, 170 1648-54.

Healy D. Pharmageddon. California University Press, Berkeley (2012).

Healy D, The Shipwreck of the Singular; Healthcare’s Castaways. Samizdat Press, Toronto (Forthcoming 2020).

Hill A.B. Reflections on the Controlled Trial. Annals Rheum Disease 1966; 25, 107-113.

Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Quarterly 2004; 82; 661-687.

Lasagna L. Discovering adverse drug reactions. JAMA 1983; 249: 2224-5.

Mangin D, Sweeney K, Heath I. Preventive health care in elderly people needs rethinking. BMJ 2007, 335 285-287.

Silberzahn S, Uhlmann EL, Martin DP, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. 2018; 1:337-56.