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

**David Healy – September 2020**

**Executive Summary**

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

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

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

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

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

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

**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, there were questions about the epistemological link between RCTs and clinical reality. Since then, there have been 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, but this questioning assumes RCTs connect with clinical reality and the only task is one of smoothing some statistical edges.

Repeated characterizations of RCTs as offering gold-standard evidence likely leave most clinicians thinking these trials have a solid epistemological foundation, even as clinicians recognize difficulties in translating from population or average effects to individual patients. In legal settings, RCTs are pitched as generating evidence that is generalizable and knowledge that lies within confidence limits in contrast to the views of clinicians and case reports. The historical record gives the lie to these views.

**Pre 1962: Hill, Fisher and Randomization**

A Medical Research Council trial of streptomycin in 1947 demonstrated the feasibility of randomization as a control of the subtle biases involved in evaluating a medicine. This trial missed the tolerance that develops to streptomycin and the deafness and other problems it causes - information evident in a prior trial of streptomycin that controlled for confounders in the then standard way and depended on clinical judgement (Healy 2012; Healy 2020).

Tony Hill, the MRC trial lead, got the idea of randomization from a horticultural thought experiment about fertilizers outlined two decades previously in which Ronald Fisher proposed that randomization could control for unknown confounders. Hill thought that randomization might control for the difficult to detect ways in which clinicians steer patients likely to respond well into an active treatment arm. Hill’s randomization was a method for fair allocation, not a means of controlling for the unknowns linked to doctors not knowing what they were doing (Healy 2020).

RCTs brought statistical significance in their wake because Fisher argued that the only things that can interfere with expert judgement not being correct every time are unknown confounders and chance. Significance testing could control for chance and randomization for unknown confounders. Fisher’s model had an anchor in the real world – an expert 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.

The more doctors know what they are doing, the more they approach Fisher’s expert, but no one runs RCTs 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.

In contrast, in trials comparing stents to other cardiac procedures, doing what seems physiologically 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 trials demonstrate 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 and for stents in some quarters prior to any RCT.

**Pre 1962: Neyman and Confidence Intervals**

Jerzy Neyman and Egon Pearson took issue with Fisher’s real-world anchor - a semi-infallible expert. They 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 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. 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.

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 during (but not 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.

In cases like this, 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. A mechanism to decide whether there are one or two stars present should not turn up the answer there are none. Distinguishing between Average Treatment Effects (ATE) and Heterogenous Treatment Effects (THE) recognizes (Kravitz et al 2004) 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 heterogenous responses are common the notion of ATE 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.

**1962: Lasagna, Hill & Primary Endpoints**

The 1938 U.S. Food, Drugs and Cosmetics Act required pharmaceutical companies to establish the safety of their products. The birth defects thalidomide caused produced a political crisis in which something needed to be seen to be done. Louis Lasagna, through Estes Kefauver, proposed that companies should also be required to demonstrate treatment effectiveness – an ineffective treatment cannot be safe. The 1962 Amendments to the 1938 Act paired the word Effectiveness with Safety throughout, with two placebo controlled RCTs later proposed by Lasagna as the means of demonstrating effectiveness.

These provisions were put in place before it was realized that demonstrating effectiveness rather than a treatment effect was not a realistic gateway to the market. In 1962, it was also assumed that RCTs offered generalizable knowledge and a positive result would invariably be replicated but this has not been borne out.

Before 1962, RCTs were not seen as offering gold standard knowledge about what drugs do. As Tony Hill put it in a 1965 lecture, 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 is mad (Hill 1965). Hill’s lecture ties RCTs to the investigation of one effect (linked to efficacy) and places the information they yield within the framework of clinical judgement.

Fisher’s significance testing, and Gauss’ confidence intervals, require a focus on a primary endpoint. In medical RCTs, a focus on a primary endpoint is key to ensuring that only chance or measurement error will get in way of the correct result. Ipso facto, this means RCTs are not a good way to evaluate a 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 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.

It is often assumed 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 the RCT measuring process is often not trained on the 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.

The measuring attention given to a primary endpoint essentially creates an act of hypnosis in which common treatment effects can be missed entirely or given an incidental status. Casting RCTs as offering gold standard evidence about a drug, rather than one effect of the drug, creates an ignorance about the ignorance they generate.

RCT evidence should never trump an evident safety 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 an effect like suicidality (and there have been none), RCT evidence is irrelevant and it is pernicious to pitch irrelevant RCTs as science that should count for more than “anecdotes” containing CDR, dose response, and other evidence.

The transformation of RCTs from a hurdle industry had to surmount to make gold into gold standard knowledge has made RCTs a gold standard way to hide the 99 other effects drugs have.

**Post 1962: Confounding & Causality**

Discussions of the results of epidemiological studies apparently linking drugs to treatment effects often caution that confounding by indication undercuts any easy assumption of a link. RCTs, which are essentially epidemiological studies, rarely come with this rider. Many clinicians likely think that randomization takes care of confounding by all unknown unknowns, including by indication, with many saying RCTs demonstrate cause and effect where other epidemiological studies produce correlations.

Consider 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 praising imipramine’s benefits also noted it could cause agitation and suicidality in some patients that cleared when the drug was stopped and reappeared when restarted. This Challenge-Dechallenge-Rechallenge (CDR) evidence, especially as it was replicated by several clinicians with different patients, offers close to Fisherian expert like certainty that imipramine causes suicide in certain individuals.

Despite being able to cause suicide, in an RCT of melancholic patients, 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, weaker than imipramine, SSRIs 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 when 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 ordinarily do not have a primary endpoint. In these, treatment effects stand out more clearly. SSRI Drug Trials in the 1980s 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. The non-confidential Drug Trial data remain 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. The depressive features IBDD patients have mean that they can readily meet criteria for MDD.

In April 2006, GlaxoSmithKline (GSK) released RCT data showing a worrying increase in suicidal events in MDD patients 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/Patients  IBDD Trials Acts/Patients  Combined 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 but it must apply to some extent in every trial that recruits patients who have a superficially similar but in fact heterogenous conditions such as depression, pain, breast cancer, Parkinson’s disease, diabetes or almost any medical disorder. 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, mask the beneficial treatment effects of an antibiotic on back pains linked to infections (up to 10% of back pains).

This is Heterogeneity of Treatment Conditions (HTC) rather than HTE. In epidemiological studies confounding by indication is commonly taken to mean that we should not for example interpret results apparently associating a treatment like an antidepressant with suicidality given the possibility that depression can cause suicidality, but in fact this effect likely works more commonly to hide the adverse effects of treatment. It is even possible to design Treatment Trials to hide adverse effects – as above.

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 a backpain RCT. We do not know enough about placebo responses to know the extent to which, in the context of randomization, they might 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 those for whom it works a placebo may be preferable to therapeutic poisoning.

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

Finally, the suicidality, sexual dysfunction, agitation, and insomnia antidepressants cause 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.

**Post 1980: From Therapeutic Poisoning to Sacraments**

In 1947, treatment with medicines was viewed as therapeutic poisoning. As of 1951, FDA made most new medicines prescription-only on the basis that they are unavoidably risky. Since the mid-1990s, however, regulators license drugs on the basis of a favorable risk-benefit profile, which implies a balance in which 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.

In 1959, clinicians could readily distinguish between treatment emergent suicidality and suicidality caused by melancholia. In 1961, Frank Ayd, the discoverer of amitriptyline a year before could distinguish the sexual dysfunction it causes from the sexual dysfunction melancholia causes. Through to 1991, clinical knowledge of the range of effects drugs can cause derived primarily from clinical experience, embodied in case reports and published in clinical journals. A 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, leading journals stopped taking anecdotes about “side” effects that almost by definition must be rare compared to the treatment effect.

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 it could take decades for significant harms such as with impulse control disorders on dopamine agonists, persistent sexual dysfunction on isotretinoin, antibiotics, finasteride and other drugs, mental disorders on fluoroquinolones or leukotriene antagonists, or dependence on psychotropic drugs, to be accepted.

This growing delay underpins a perception that pharmacovigilance is in crisis. Proposed solutions mention the need for systems to detect rare treatment effects 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 one effect 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. Nor can the data-handling methods developed for fertilizers and stars be expected to encompass the complexity of therapeutic poisoning.

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, and one patient may have both kinds. Instruments (checklists) specifically designed with the characteristics of each star in mind may facilitate the distinction in research or practice, but ultimately in practice 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 operational approach appealing – but it is not good science or good medicine.

If registered on adverse event forms, treatment emergent suicidality or sexual dysfunction should almost de facto be causally linked to treatment. Without clinical context and the opportunity to dechallenge and rechallenge, faced with a requirement to tick boxes as to the likelihood of a link, the ethos of RCTs, which replaces clinical judgement with decisions based on analytic processes rather than an interrogation of people, steers investigators toward possibly unrelated. As a result, treatments are transformed from poisons into sacraments – (hyperreal) agents from which only good can come.

Facing claims in 1983 that spontaneous reporting of adverse events was unsophisticated and not scientifically rigorous, and the only proper method of establishing effects was through trials, Lasagna, once a leading advocate for RCTs, responded that “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 [trials]” (Lasagna 1983).

**Implications: Objectivity**   
  
A few years later, Lasagna offered the view that: “In contrast to my role in the 1950s which was trying to convince people to do controlled trials, now I find myself telling people that it’s not the only way to truth… Evidence Based Medicine has become synonymous with RCTs even though such trials invariably fail to tell the physician what he or she wants to know which is which drug is best for Mr Jones or Ms Smith – not what happens to a non-existent average person”.

Concerns about what is often termed the population effects of RCTs, or Average Treatment Effects (ATE) and the mismatch between these and the 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. Designating RCTs as offering gold standard evidence, however, effectively side-lines and denigrates the judgements of clinicians and patients.

The view that RCTs give population or average treatment effects assumes a valid population with individual outliers. In the case of antidepressants, however, there is no knowing how ***any*** individual will respond. 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, and individual heterogeneity we do not have valid populations. To adapt Shakespeare, the fault lies in our stars not in ourselves.

In recent years, there has been sophisticated consideration of the statistical techniques employed in epidemiological studies, including RCTs (Greenland et al 2017), and of the merits of RCTs applied to other complex situations in the social sciences (Deaton and Cartwright 2019). Both considerations have stressed the role of judgement in deciding what populations and experimental design are appropriate, and how results should be interpreted. Both view RCTs, and related designs using statistics, as assay systems yielding results specific to the system, rather than experiments that generate ‘knowledge from nowhere’ – we don’t have to worry whether the laws of gravity will apply to the next patient.

These positions are compatible with the argument here, which is that rather than assay systems that might in the right circumstances offer applicable information, RCTs have become algorithmic or operational procedures. 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, 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 supposedly 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 (data) that challenge prior judgements. The application of statistical techniques to data yields outputs, not observations. While these techniques and their outputs can be useful, the mission of science has not been to replace judgement by technical outputs.

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 Mayo Clinic streptomycin trial demonstrated; the idea that RCTs as happen now in clinical settings have as clear a footing is assumed not established.

Arguments favoring RCTs to 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 internal mammary ligation trial only happened because the dominant clinical judgement was that this treatment didn’t work – an article in the Reader’s Digest notwithstanding. Randomization didn’t work in this 17-patient trial.

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.

Our most important failure is our complicity in a sequestration of trial data, fooled perhaps in some instances into thinking that analytic outputs are data. Data means the people entering into a study, who lie behind any table of figures or the outputs of any analytic process applied to those figures. At present, with the exception of a very few RCTs, case reports with names attached are the only form of controlled clinical investigation that offer access to and the possibility of interrogating the data and, accordingly, an opportunity to ground any conclusions in the real world.

Drug interventions (therapeutic poisoning) invariably harm; the hope is that some good can also be brought from their use. Evaluations of a medicine 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 et al. 2018).

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 persons (interrogate the data) given that treatment. 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.

**Implications: The Place for Randomized Controlled Trials**

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.

An increasing use of RCTs in social science, economic and political settings makes it clear that complexity is not a necessary bar to their use in trials with an appropriate focus on a primary endpoint. In medicine, the multidimensional nature of therapeutic poisoning adds an extra layer of complexity and makes a focus on a primary endpoint problematic, other than when a claimed benefit is contested.

There is a place for RCTs 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 an 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 prefer to die by 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, although strictly speaking this is inconsistent with current statutes.

After 1962, RCTs became the standard through which industry would make gold, As they proliferated, 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.

The launch of a drug licensed on the basis of a treatment effect would be the point when more comprehensive clinical evaluations should start, aimed at generating 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, and to the values of patients increasingly need to reduce their medication burden to achieve optimal benefits.

In addition, 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, which need to be distinguished from superficially similar condition effects.

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 trial data and ghostwriting 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. 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.

The incorporation of RCTs into the regulatory apparatus has introduced 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, have more 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 based on the outputs from analytic processes rather than in clinical reality.

Possible answers to these problems lie with medical journals who should insist on the data from Drug and Treatment Trials being published. Our hierarchies of evidence should come clean on whether they regard a ghostwritten article without access to 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 might translate into the real world, rather than assuming they do.

But these problems avail of an epistemological gap between the conduct of an assay and a consideration of the implication of any results rather than constitute the gap.

**Coda**

Evaluating treatment effects 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 topped by 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. Succumbing to an operational solution, however, is at least as dangerous as depending on judgement.

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). Adding parachutes and gadgets that are effective (rather than just have an effect) enhances aviation safety, although recent Boeing crashes point to the perils of too great a reliance on automatic decision tools. Combining five pluripotent drug gadgets almost certainly brings risks of interactions that airplane gadgets don’t bring and current data indicates that reducing medication burden from 10 or more drugs to 5 or less reduces hospitalization, increases life expectancy and improves quality of life (Garfinkel and Mangin 2010). If RCTs of medicines essentially produce evidence that it is not correct to say this drug has no possible benefit, rather than that they are effective, the prospect that our treatments might be a contributing factor to increasing levels of mortality and morbidity would make more sense.

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). So does data indicating antidepressants are now the second most commonly used drugs by young women in the face of 30 out of 30 trials negative on the primary outcome, which advocates of RCTs, with no links to industry, claim to be able meta-analyze and extract positive effects from data taken from ghostwritten publications, without access to trial data.

**References:**

Deaton A, Cartwright N (2018). Understanding and misunderstanding randomised controlled trials. Social Science and Medicine 2018, 210, 2-21.

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.

Greenland S, Senn SJ, Rothman K, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol, 2016, 31:337–350.

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.