Research Article

Jürgen Landes* and Daniel J. Auker-Howlett

Current philosophical perspectives on drug approval in the real world

https://doi.org/10.1515/jci-2023-0011
received March 08, 2023; accepted March 25, 2024

Abstract: The evidence-based medicine approach to causal medical inference is the dominant account among medical methodologists. Competing approaches originating in the philosophy of medicine seek to challenge this account. In order to see how successful these challenges are, we need to assess the performance of all approaches in real world medical inference. One important real world problem all approaches could be applied to is the assessment of drugs for approval by drug regulation agencies. This study assesses the success of the *status quo* against an empirical non-systematically obtained body of evidence and we scrutinise the alternative approaches from the armchair, contemplating how they would fare in the real world. We tentatively conclude that the *status quo* is regularly not successful at its primary task as it regularly fails to correctly assess effectiveness and safety and suggest that this is due to inherent factors of the "messy real world." However, while all alternatives hold promise, they are at least as susceptible to the real world issues that beset the *status quo*. We also make recommendations for changes to current drug approval procedures, identify lacunae to fill in the alternatives, and finally, call for a continuation of the development of alternative approaches to causal medical inference and recommendations for changes to current drug approval procedures.

Keywords: drug approval, causal inference, randomised controlled trials, causation, evidence synthesis

MSC 2020: 62D20, 62R07, 62A99, 60A99, 62C99, 62F15

1 Introduction

Drug licensing agencies perform a vital job. They are tasked with approving newly developed medications for market authorisation, as well as assessing the safety of currently approved medications via continued monitoring (pharmacovigilance). To carry out their job, agencies rely on a wide set of tools and procedures. A big part of the work they do is evaluating causality. Casual inference in medicine is particularly challenging since we often cannot observe the workings of the human body in real time at the level of interest.

Evidence-based medicine (EBM) was developed to support health professionals to make individual clinical decisions. In order to make good decisions, clinicians and patients need to estimate the outcomes of possible treatments as well as the patient's preferences over health outcomes (utilities). In order to estimate the outcomes of treatments, their causal effects need to be inferred. In the following, we are only interested in the epistemology of causal inference of EBM, which we denote by EBM_{epis}. With this terminology in mind, we can state that EBM_{epis} provides the dominant account of causal inference in medicine. According to this account, different kinds of evidence are better or worse at supporting the estimation of causal relationships [1,2].

*** Corresponding author: Jürgen Landes,** Department of Philosophy "Piero Martinetti", University of Milan, Milan, Italy; Munich Center for Mathematical Philosophy, Geschwister-Scholl-Platz 1, Munich, Germany, e-mail: juergen_landes@yahoo.de
Daniel J. Auker-Howlett: Independent Researcher, Cambridge, UK, e-mail: djaukerhowlett@gmail.com

Randomised controlled trials (RCTs) are the "gold standard" according to $\mathrm{EBM}_{\mathrm{epis}}$ because (1) patients are randomly allocated to different interventions (e.g. taking a drug vs taking a sugar pill), (2) one group can (but not always) be given an inactive intervention, and (3) both participants and researchers are unaware of what intervention is applied to which group – this is called "blinding." The measured difference between the states of the trial groups is modulo statistical error, taken as the difference of causal effects of the various interventions. Methodological strictures (1–3) explain why.

First, randomisation (1) facilitates blinding and helps to balance the treatment arms with respect to potential confounders. The alternative might be that only really sick patients took a drug because caregivers were appropriately attendant to their needs, and matched controls were not given the drug because they were going to get better anyway. Second, comparing an active and inactive intervention (2) rules out the possibility that the measured effect is a "placebo effect," where the process of administering treatment causes beneficial effects through the elicitation of participants' own internal self-healing processes. Finally, (3) ensures that the effect is not a result of preferential treatment by researchers or elicitation of the placebo effect through participants' knowledge of which group they are in. Clearly, leaving out these methodological strictures may lead to inaccurate causal inferences. Note, however, that it is not always possible to achieve the lofty gold standard. For example, it is impossible to implement blinding in comparative trials of behaviour therapy to a psychoactive drug or in trials comparing surgery to physiotherapeutic interventions. Nevertheless, if either of (1–3) blinding is achievable in a context, then it should be implemented for that evidence to be considered gold standard.

Philosophers of science have criticised the $\mathrm{EBM}_{\mathrm{epis}}$ paradigm for some time [3–8]. The main critique is aimed at $\mathrm{EBM}_{\mathrm{epis}}$'s fixation on evidence from experimental studies that employ randomisation at the expense of evidence obtained from other sources, e.g. observational studies without randomisation, laboratory studies. Some philosophers went beyond criticising $\mathrm{EBM}_{\mathrm{epis}}$ and developed alternative approaches to causal medical inference. Each alternative advances predominantly philosophical reasons for why their approach improves on EBM, and they all agree on putting more emphasis on evidence obtained by means other than RCTs. For example, the motivation for EBM+ is that its logic of causal evaluation is more secure than $\mathrm{EBM}_{\mathrm{epis}}$'s (see Section 3.1 for details on EBM+). This approach has been criticised as using "friction-free epistemology" [9] (see also [10,11]), in that it ignores the factors relevant to human actors and their various biases. We think that while one need not worry too much about human biases when weighing the relative merits of one logic of causal evaluation over another, one should worry when evaluating applications of causal inference methods to problems such as drug approval.

This study assesses the success of the "*status quo*" of drug approval and its alternatives. The *status quo* utilises $\mathrm{EBM}_{\mathrm{epis}}$-like causal inference methodology.[1] We argue based on an empirical non-systematically obtained body of evidence that the *status quo* has demonstrable problems with establishing effectiveness[2] and safety. Furthermore, the novel causal inference methods, developed in a friction-free approach to epistemology/philosophy of science, would plausibly fare badly when applied to drug approval. Andreoletti and Teira [14] presented a very useful first step in this evaluation endeavour, raising the problem, presenting an assessment criterion, and carrying out a preliminary assessment (see also [15]). We build on this and suggest avenues for future research.

The rest of this article is organised as follows. We next delineate assessment criteria and assess the *status quo*. A path to a more complete assessment is given in Section 2.5. The alternative approaches are considered in Section 3. Section 4 concludes, reflects, and recommends.

---

1 While there are some differences between $\mathrm{EBM}_{\mathrm{epis}}$ and the causal account used by drug regulators, we shall here not distinguish between them, since both (i) strongly emphasise randomised study designs for assessing effectiveness at the expense of other forms of evidence, (ii) heavy rely on statistical techniques, (iii) are strongly influenced by [12], and (iv) agree on a somewhat increased importance of other forms of evidence for safety assessments, cf. [11,13]. Given the number of people involved on both sides, there is some unavoidable vagueness at the boundaries and possible incremental shifts at any time.

2 *Effectiveness* and *efficacy* are related yet different concepts: the former concerns the intended effects in the real world, while the latter refers to intended effects under controlled circumstances.

2 Assessment of the *status quo*

2.1 Assessment criteria

Drug approval decisions present a choice between granting and refusing approval. The quality of decisions can be measured by how well we are then able to treat patients and at which costs.[3] How well we are able to treat patients is determined by trading off the *effectiveness* and *safety* of a drug.

As mistakes are likely to be made, it is imperative to quickly identify and rectify errors. The measure of success of a drug approval procedure is therefore determined by the rates and seriousness of errors in determining effectiveness and safety as well as the time it takes to discover poor decisions.

In what follows, we evaluate whether the *status quo* is successful against three criteria:

A1 How often are safe and effective (and cost-effective) drugs not granted approval because they are assessed as unsafe or non-effective (or not cost-effective)?

A2 How often are unsafe or non-effective (or not cost-effective) drugs granted approval because they are assessed to be safe and effective (and cost-effective)?

A3 How swiftly are errors corrected?

The first two criteria capture the possible directions of error (falsely rejecting [A1], falsely approving [A2]) while the third criterion captures the time it takes to correct errors. Overall, the criteria plausibly capture the quality of the *status quo* as a decision making process.

A1 and **A2** require refinements. The properties "safe," "effective," and "cost-effective" are, of course, not binary properties. Drugs are effective, safe, and/or cost-effective to different degrees. This variation depends on a number of other factors, such as severity of the condition to be treated; other available treatment options; and how well possible adverse reactions can be treated. Furthermore, not only does it matter how often an approach gets it wrong (its error rates), the severity of the error matters too. Approving an ineffective drug causing only, few and mild adverse reactions is preferable to approving an ineffective drug, which often causes severe adverse reactions. In the following, we have these more refined criteria in mind.

We distinguish between data (observations/measurements), evidence (data and information on the reliability of methodology, instrumentation and experimenters), and confirmation (how strongly evidence confirms hypotheses). The degree to which evidence confirms hypotheses depends on assessed reliabilities, assessed biases, and a confirmation method (intuitive, Bayesian, frequentist, or others). Evidence thus only confirms via an interpretative act. We shall write "we think that" (or similar) to signal that we perform an interpretative act when assessing confirmation of a body of evidence.

Our assessments of effectiveness and safety for the *status quo* are based on a body of evidence that has not been systematically obtained nor do we carry out a quantitative analysis. Both these tasks are outside the scope of a philosophy study. Section 2.5 comments on how to carry these tasks out in principle.

2.2 Evaluating effectiveness

In this section, we describe the *status quo* for evaluating effectiveness during the drug approval process and then assess whether that process succeeds against our assessment criteria.

3 We here take the term costs to include drug development costs, drug prices, but also costs arising from the training of medical researchers, prices of drug approvals, and so on. While we do think that such costs are a relevant assessment criterion, we will have preciously little to say about them in this article.

2.2.1 Approval process

In the European Union (EU), the European Medicines Agency (EMA) advises the European Commission on drug regulation based on assessments of evidence. At the time at which an assessment regarding immediate market entry is made, EMA focuses on (typically two phase III) RCTs[4] which supposedly demonstrate beneficial effects. The possibility of adverse reactions caused by the drug is assessed based on these RCTs as well as on other evidence (such as phase 0 to II trials, animal studies, and computer modelling). We here focus on the standard decision procedures, i.e., we are not considering (i) fast track drug approval for serious conditions nor (ii) orphan drugs intended for so few patients that RCTs with sufficiently many patients to achieve an adequate level of statistical power become infeasible.

The Food and Drug Administration (FDA) in the USA used to follow similar decision procedures [16]. The 21st Century Cures Act in the USA, signed into law in 2016, now also allows applicants of market authorisation to present, and the FDA to rely much more on real world evidence (RWE). RWE can refer "*to information on health care that is derived from multiple sources outside typical clinical research settings, including electronic health records (EHRs), claims and billing data, product and disease registries, and data gathered through personal devices and health applications*" [17].

In essence, the European approach, according to EBM_{epis}'s heavy emphasis of randomised studies is to subject drugs that hold promise based on previously accumulated evidence to be effective and reasonably safe to two RCTs. In case the RCTs confirm this initial effectiveness assessment (and the drugs are assessed as safe enough; see Section 2.3.1), EMA recommends approval,[5] which is then granted by the European Commission.

The approach was developed to put an end to exaggerated claims of benefit [14]. It was thus designed with effectiveness assessments in mind. Beneficial effects, if they do obtain, obtain within a well-understood time frame and well-defined parts of the human body. RCTs are an appropriate means to learn about effectiveness: RCTs are expensive[6] and thus, study not too long time frames; RCTs analyse the specific parts of (patho-) physiology and lend themselves to statistical (meta-)analysis to detect significant effects.

However, practical problems abound even when the issues concerning randomisation mentioned in the introduction can be avoided. First, defining a statistical measure prior to conducting RCTs is not a trivial matter (progressing from determining a target of estimation [estimand], via fixing methods of estimation [estimators] to calculating numerical results [estimate] including sensitivity analyses) [18]. Second, some planned measurements of variables may not occur (e.g. a patient switching to a different treatment or dying) or may be difficult to interpret (e.g. a patient using an additional medication made available for ethical reasons) [18]. Third, there is the hard methodological problem of passing from an efficacy estimation to effectiveness assessments [19–22].

Until recently, reliance on two RCTs was also the essence of the approach in the USA. Not only did the EU and the USA implement similar decision procedures, they also often made the same decision [23]. Given the provisions in the 21st Century Cures Act, these similarities are set to become less pronounced.[7] To date, there is limited scholarly research on the effect of these provisions on drug regulation [24]. We hence first focus on EMA's current approach and then briefly consider changes brought about by the 21st Century Cures Act. EMA's current approach will be referred to as the *status quo*.

---

4 Phases are numbered chronologically. The first three phases are typically pre-approval and later phase trials are usually larger; phase IV are post-marketing trials [16].

5 Other evidence from the pre-approval phase is not ignored but does play second fiddle to the randomised studies. Without positively evaluated RCTs, there are no approvals.

6 "*Shortening time from concept to market is not only important to patients. During that period in a drug or device's life, it generates costs rather than revenue for its sponsor. For drugs, most of that time, in both Europe and the United States, is spent in clinical trials that can consume years and generate costs in the millions or even billions of dollars*" [16].

7 As we say in Section 2.5, we believe that there will be only a few historic cases in which drug regulation agencies have come to different decisions, although this may change in the future.

2.2.2 The evidence

How accurately does EMA's approach assess the effectiveness of drugs? We now present evidence that drug regulation agencies regularly approve drugs with disappointingly low effectiveness.

"By law, the German health technology assessment agency IQWiG (Institute for Quality and Efficiency in Health Care) must investigate the added benefit of new drugs compared with standard care" [25] (the methodology used for these assessment is available on IQWiG's website [26]). IQWiG's findings are stark:

> Only 54 of the 216 assessed drugs (25%) were judged to have a considerable or major added benefit. In 35 (16%), the added benefit was either minor or could not be quantified. For 125 drugs (58%), the available evidence did not prove an added benefit over standard care for mortality, morbidity, or health related quality of life in the approved patient population. [25]

The same review found that for new drugs for over 100 indications approved by the FDA

> superior efficacy on clinical outcomes was confirmed in less than 10% of cases. A higher but still insufficient rate (20%) was shown in a similar publication on cancer drugs. [25]

Wieseler et al's findings are widely supported. The proportion of new approved drugs that actually improved on existing pharmaceuticals in the real world ranges between 10 and 20% in a number of reviews that each investigated hundreds of cases [27–30].

These findings are widely acknowledged, yet seem to make little impact on methodology. For example, a paper on pain medication with the title: "The Development of New Analgesics Over the Past 50 Years: A Lack of Real Breakthrough Drugs" concludes starkly:

> Morphine and aspirin, introduced for the treatment of pain more than a century ago, continue to dominate biomedical publications despite their limited effectiveness in many areas (e.g. neuropathic pain) and multiple serious adverse effects. [31]

Furthermore, a recent editorial in the BMJ is aptly subtitled: "*We must raise the bar to ensure real benefits for patients*" [32]. We agree – many drugs are being approved with low effectiveness.

Given the aforementioned body of non-systematically obtained evidence mainly comparing the effectiveness of new drugs to the effectiveness of older drugs, we believe that the *status quo* approves non-effective drugs because those drugs are assessed as efficacious: the *status quo* regularly approves drugs with disappointingly low effectiveness, we here mean absolute and not comparative effectiveness. This is in the interest of drug manufacturers.

We think that changes to the *status quo* are required to improve the effectiveness of approved drugs (Section 4).

In the USA, the 21st Century Cures Act also allows drug manufactures to submit RWE supporting their applications for drug approval. Such evidence, which can be presented as the applicants see fit, can only lead to further overestimation of effectiveness. Scholarship on this act is still in its infancy [33,34], but there are the beginnings of some relevant empirical evidence:

> We are concerned that the use of RWE for regulatory drug approval will increase uncertainty over the true benefits of therapies we offer to patients. As investigators who work routinely with RWD (Real World Data), we have concerns that two of the highlighted FDA decisions were based on RWE that came from case series of only 14 and 26 patients. Our patients deserve timely access to new cancer therapies; however, they also deserve therapies for which we have strong evidence showing meaningful gains in quality and/or quantity of life. [24]

2.3 Evaluating safety

Small effect sizes do not always make a drug useless. A drug with a small effect size may confer a net benefit to the patient if its beneficial effects outweigh its harms. A drug with a small effect, yet even less harms, is clearly

beneficial in the sense of effectiveness-harm trade-offs. So, we must distinguish between effectiveness and net benefit. The preceding analysis argued that the *status quo* regularly approves drugs with disappointingly low effectiveness. However, it could still be the case that the harm profile of those same drugs is such that an effectiveness-harm trade-off would conclude that they were beneficial overall. In this section, we present evidence that shows the *status quo* does not adequately assess the safety of drugs. Therefore, the process approves unsafe drugs as safe, which means it does badly against our criteria, and we also cannot reliably conclude that drugs with small effects confer a net benefit, strengthening our arguments about effectiveness.

2.3.1 Approval process

Unlike intended effects, adverse drug reactions (ADRs) can occur after a long time, e.g. years of treatment with olanzapine causes tardive dyskinesia [35], treatment of pregnant women with diethylstilbestrol caused vaginal adenocarcinoma in their pubertal and adult children [36]. ADRs also occur in unpredictable places and may be rare yet deadly (in some cases, 1 fatality in every 10,000 patients [37]). RCTs are typically too short – "*often months to 1 or 2 years*" [38] – and too small – "*often dozens to a few hundred*" patients [38]) – to properly inform harm assessments [39–42]). In light of this, EMA, in accordance with EBM_{epis}, bases safety assessments on not only the data from two RCTs submitted for effectiveness assessments, but also emphasises other available evidence. Indeed, the EU is now calling for efforts to amalgamate safety signals such as spontaneous case reports, data-mining, pharmacoepidemiological studies, drug utilisation studies, and non-clinical studies (Directive 2010/84/EU; Regulation (EU) No 1235/2010).

2.3.2 Evidence

Andreoletti and Teira [14 emphasis original] claimed that "market withdrawals provide an *empirical benchmark*" for assessing approaches. There are several aspects to withdrawal of approved drugs. One simple measure is the number of drugs withdrawn. We agree that determining the withdrawal rate is a sensible and feasible benchmark. It helps us to understand how often bad drugs are falsely granted approval. However, as Andreoletti and Teira note [14]:

> Number of drug withdrawals is admittedly a rough index of regulatory success. For instance, there are no universal guidelines, and therefore, there is no perfect international agreement about which drugs should be available [43]. With every caveat in mind, less than 2 percent of new drug approvals by the FDA between 1950 and 2011 were withdrawn [44].

One might think that withdrawal of only 2% of drugs indicates a strong approval process. However, there are reasons to think the reported withdrawal rate is underestimated. For one, drug licensing agencies have recently begun performing paid consulting work for drug manufacturers [45]. As this is a new development, it is likely not to have impacted the numbers of withdrawals of already approved drugs reported in the study of Onakpoya et al. [44].

A stronger reason is that there are cases in which a drug was not withdrawn, but weaker regulatory decisions were taken, e.g. a change of labelling information concerning ADRs. EMA's current procedures for monitoring effectiveness and safety are described in the study of Brown et al. [46], who "identified and reviewed all EMA post-market approval referrals made for safety and/or efficacy concerns which were evaluated by an assessment committee between 1 January 2013 and 30 June 2017." The authors found that 83% of referrals led to changes to product information, while 23% led to suspension of withdrawal of market authorisation. These results are echoed in studies on FDA-approved pharmaceuticals: from 2001 to 2010, 32% of novel therapeutics were subject to post-market safety events (withdrawals, boxed warnings, or safety communications) [47]. While withdrawals were rare (3/222), new boxed warnings were not (43/222) – boxed warnings indicate potentially life-threatening or preventable safety events. Last but far from least, there is evidence that a larger share of novel cancer drugs reduce patient safety rather than improve it (45% compared to 15%) [48]. Hence, the *status quo* is making errors that have *severe* consequences.

This evidence of continued updating of safety labelling shows that (i) serious pharmacovigilance efforts continue to take place; (ii) unsuspected ADRs continue to emerge, and estimated frequencies of ADRs continue to rise, after approval; and (iii) approved drugs are actually less safe than deemed at the approval stage. One then asks whether these drugs should have been approved in the first place, and whether they should be withdrawn rather than re-labelled.

Furthermore, regulatory actions – in particular withdrawals – occurring at the first sign of trouble after approval make us wonder about the robustness of the approval and withdrawal processes. How can a few case reports alter the evidence base so drastically that regulatory action becomes necessary? This concern is echoed in [49]:

> [T]he speed with which...adverse reactions [appear] in the literature following launch...arouses suspicion of lack of transparency in reporting of harms in clinical trials conducted in the pre-approval phases or flaws in the ways in which harms [are] assessed by regulators. This is also supported by the speed with which...analgesics [are] withdrawn from the market following the reports of such reactions. Indeed one such analgesic, pifoxime, was withdrawn in 1976, within 3 months of regulatory approval in France, following reports of serious neuropsychiatric adverse reactions.

It is plausible then that many drugs that should be withdrawn are being kept on the market through changing of labelling information. At the very least, it is clear that many drugs are approved that in actual fact have potentially severe ADRs not picked up at the approval stage.

Recall that it also matters how quickly issues are rectified (**A3**). Unfortunately, for the *status quo*, it looks like it does not do well on this count either:

> We found 95 drugs for which death was documented as a reason for withdrawal between 1950 and 2013. All were withdrawn in at least one country, but at least 16 remained on the market in some countries. ... However, in 47% of cases more than 2 years elapsed between the first report of a death and withdrawal of the drug, and the interval between the first report of a death attributed to a medicinal product and eventual withdrawal of the product has not improved over the last 60 years. [50]

Pharmacovigilance was not a thing 65 years ago [43,51]. This evidence strongly suggests that harmful drugs are often withdrawn much too late. As noted, for all our benefit, some of these drugs should not have been approved in the first place.

Overall, we think that the evidence demonstrates that there is underestimation of ADRs by the *status quo* beyond the reported 2% drug withdrawal rate. Accordingly, we find that this approach, implemented in the real world, regularly approves drugs that are disappointingly unsafe: there are a large number of ADRs post-approval; the severity of errors is large; the timeliness of withdrawal is inadequate.

We are not aware of articles systematically reporting on the (un-)safty of drugs approved under the 21 Century Cures Act. A previous law change to allow for faster drug approval by the FDA was found to correlate with a decrease in safety of approved drugs [52]. We have no reason to believe that the more recent law change improves safety assessments.

2.4 Success of the *status quo*

According to the non-systematically obtained empirical evidence, we have presented the *status quo* does badly against our criteria. Given that a drug approval process should make the right calls on effectiveness and safety, and swiftly rectify arising issues, we conclude that the *status quo* is regularly unsuccessful and should be amended. In our analysis, the *status quo* receives a worse grade than in the analysis of Andreoletti and Teira [14]. They did not consider effectiveness and only considered the drug withdrawal rate as a benchmark for un-safety. Our analysis also showed the *status quo* performs poorly when judged against effectiveness and further safety criteria.

One explanation for the *status quo*'s problems is a frequent lack of a large evidence base at the time of first drug approval. Because access to innovative drugs is important, it is argued that speed of approval is worth prioritising [25]. A consequence is that studies may not last long enough to capture all harms. It thus seems natural to require the drug regulators mandate post-marketing studies. At times, they indeed do. What happens then?

> Despite their promise, a critical and well known problem with post-marketing studies is they often do not happen. Analyses have found that only about half were completed on time or within five to six years. [25]

This is not an isolated problem: post-approval trials confirm the superior efficacy demonstrated in pre-approval trials in less than a third of FDA-approved new drug indications [53]. Moreover, post-approval studies can remain incomplete for years, even when they are required by FDA and EMA [54,55]. [56] put it bluntly

> The era of post-approval (industry-sponsored) prospective observational studies seems to be mostly over.

This cuts across the benefit/harms distinction: low effectiveness in the presence of lower harm may result in an overall benefit, but to even begin to make this calculation, we first need accurate estimation of effectiveness.

Unfortunately, the failure to conduct post-marketing studies seems to have very limited consequences. Drug regulators could, in principle, grant drug approval conditional on the realisation of well-conducted and properly analysed post-marketing studies that also demonstrate a benefit (in terms of benefits, adverse reactions, and/or costs). As a matter of fact, they do not. The question arises whether the failure to grant such a conditional approval of a drug is a flaw that is intricately connected to the current regime or whether it is a modifiable ancillary feature. We lean towards the latter view. However, it is not something that requires methodological tinkering to solve. This is a political problem – the agencies can only do so much as their mandate allows. This is an example of a real world problem that goes beyond the logic of a method for causal inference. Accordingly, the solutions are external.

Another explanation starts with recognising the issues caused by the combination of the "2 trial requirement" and the fact that the *status quo* puts the burden of proof on the drug manufacturer seeking market approval. As such, the manufacturer generates, collects, and presents evidence to the drug licensing agency and so the manufacturer chooses how evidence is presented. The problem is that, in principle, a manufacturer can run a large number of trials and present the most favourable of them to the drug licensing agency. Relative to this large number of trials, natural variations or random noise can plausibly lead to two trials, which clearly document the efficacy of a drug. The entire body of evidence may, of course, entail different conclusions.

This is not a hypothetical issue. The drug manufacturer Roche ran (at least) **107** trials studying the drug Tamiflu (oseltamivir). Tamiflu was approved by EMA on the basis of ten trials (BMJ). It was also stockpiled around the world to combat the H5N1 pandemic (bird flu): governments spent billions to purchase this drug. A systematic review of these 107 trials (obtained after a prolonged struggle with Roche) instead concluded that the purported benefits did not outweigh the known harms. Therefore, the review questioned "the stockpiling of oseltamivir, its inclusion on the WHO list of essential drugs, and its use in clinical practice as an anti-influenza drug" [57].

As a result, EMA now has a policy to publish all trial data contained in applications for market authorisation [58].

As far as we are aware, no Roche employee has been charged in connection with the Tamiflu debacle. Roche has not been fined for its actions. Only now are legal efforts beginning to attempt to claim back the billions of dollars governments spend on stockpiling Tamiflu (see, e.g. [59], https://www.prnewswire.com/news-releases/judge-rules-tamiflu-maker-hoffman-la-roche-must-answer-whistleblower-fraud-claims-301144247.html and https://casetext.com/case/united-states-v-roche-holding-ag). For more background and details on the Tamiflu debacle, see the seminal BMJ campaign, and also [58] for a more philosophical take. Methodologically, the requirement of assessing the *total* evidence is well known [60, Section 3] and widely accepted. If this principle had been adhered to in the Tamiflu case, the 97 trials excluded should have factored into decision making. It is reasonable to think this would be an important principle to adhere to in general.

Sponsorship bias and too short follow-up periods in RCTs [61], poorly handled inductive risks in the post-approval stage [62], failure to report on harms [63] as well as underpowered RCTs (Section 2.3.1) (see also [25]) are further possible explanations. Poor oversight by regulators also plays a role [64]. Together, these issues are all part of the messy real world and plausibly are not solvable by improvements in our tools for doing causal inference. We are hence less positive than [14] about the *status quo*. We think that changes to the *status quo* are required (Section 4).

2.5 Operationalisation of assessment criteria

A defender of the *status quo* may argue in turn that our examples are cherry picked, present the worst of a small set of mistakes, and that in the main, the *status quo* is successful. Indeed, and though they may not be such defenders, Andreoletti and Teira argue that we should in some sense *measure* how well an approach scores according to some *operationalised* criteria [14]. Although it is beyond the scope of this article to carry out a full, quantitative assessment of the drug approval landscape, we next operationalise our assessment criteria. The point of this is two-fold: (1) to show how our assessment might be practically applied; and, (2) to provide a jumping off point for those interested readers with the requisite skills and time to carry out a fuller evaluation. Turning these criteria into a single overall assessment is a relevant and highly complicated task, belonging to the realm of medical multi-criteria decision analysis. Acceptable error rates depend on the targeted disease, personal preferences and on whether one is a patient, regulator, or doctor [65–68]. Addressing these questions properly is outside the scope of this article – it deserves at least a full article on its own.

A1: How often are safe and effective (and cost-effective) drugs not granted approval because they are assessed as unsafe or non-effective (or not cost-effective)?

We believe that one cannot assess this error rate and the gravity of the errors without access to an all-seeing oracle. However, we see two scenarios that may be instructive. One should measure:

(i) The extent to which different drug regulators make different decisions [69], e.g., the EMA might approve a drug, while the FDA does not grant approval. If the drug later turns out to be good, then the drug was falsely denied approval by the FDA.

(ii) The number of approved good drugs A which are similar in all relevant aspects to drugs B that were previously not granted approval – we may be in a position to infer that drug B was falsely rejected. We believe that this second scenario is rare, since we will be only very rarely in a position in which we can infer that drug A is similar in all relevant aspects to drug B.

A2: How often are unsafe or non-effective (or not cost-effective) drugs granted approval because they are assessed to be safe and effective (and cost-effective)? One should measure:

(i) Withdrawal rate (see above)

(ii) *Severity* of errors can be measured by the many actions a regulator may take. Actions include withdrawals, additions of known counter-indications, as well as additions to the packaging leaflet of possible adverse reactions, their severities, and their (estimated) frequencies.

(iii) How well approved drugs perform in follow-up studies and/or comparative trials – this measures the actual benefits of such drugs. For example, a process that approves many drugs that turn out to offer no or little benefit over existing drugs would do badly against this criterion.

A3: How swiftly are errors corrected?

(i) Measure the time it takes a regulator to take action.

In this section, based on an empirically informed methodology, we have shown that it is very plausible that the *status quo* is regularly unsuccessful when it comes to drug approval when evaluated against **A2** and **A3**. An evaluation against **A1** is currently infeasible due to similarity of decision procedures (cf. Section 2.2.1) and the rarity of sufficiently similar drugs with different regulatory drug approval decisions. The lack of evidence concerning **A1** is unfortunate given the tension between **A1** and **A2**: it is possible to be perfect with respect to **A1** but completely fail at **A2** by simply approving every drug; vice versa, simple rejecting every application of drug approval will earn a perfect **A2** score but utterly fail at **A1**. Nevertheless, our analysis shows that the *status quo* is less than ideal. Our operationalisations invite a fuller quantitative assessment to further confirm our conclusions.

3 Alternative approaches

If the *status quo* is regularly unsuccessful, then we may want take a similar approach to that of a recent trend in philosophy of science: where problems have been raised with $\mathrm{EBM}_{\mathrm{epis}}$ in general, alternative approaches have been proposed. In this section, we assess three such alternatives as they might be applied to the problem of drug approval.

However, a similarly empirical assessment of the alternative approaches is unfortunately impossible – none of the approaches have been used for drug approval; hence, there are no empirical data to assess their methods against. At best, we can assess the approaches with a philosophical analysis, contemplating how they might perform at approving drugs.

In order to stay as close to the spirit of our assessment of the *status quo*, we consider plausible ramifications of situating each approach in the drug approval context where drug development and manufacturing is a multi-billion dollar industry. The enormous sums at stake necessitate that at every step the (possibly subconscious) agendas of all involved parties and stakeholders have to be considered. Whatever the decision procedures an approach puts forward, once they are in place, an (epistemic) arms race between drug developers, regulators, and other stakeholders ensues [70].

Some of the behaviour to exploit, bend, or even break rules can be predicted and modelled [71,72] and the impacts can be, to a degree, assessed. Given the track record of creative ways of influencing drug regulation, prescription, and use, such as guest-authorship for renowned medical researchers [73],[8] we cannot predict the ways in which different procedures for drug approval can and will be exploited, bent, and broken. We assume that it will happen and draw from that assumption some implications for the alternatives. We hence require an epistemology and decision-making procedures that do not abstract away the complexity of the drug approval context [9].

Before considering three alternatives in detail, we identify three more and explain why we do not analyse them.

- The CauseHealth approach aims to improve our understanding of causality in medicine from a dispositionalist point of view [90,91]. While we are sympathetic to many of their views, the real world applicability of this approach to serve as a drug approval decision procedure has not yet been achieved: too many details for performing causal inference in practice have not yet been specified.
- The Union of Soviet Socialist Republics (USSR) approach nationalises all pharmaceutical industries. [14] seem to suggest that we can use the number of drug patents to measure (the failure of) innovation in this approach. To us, this seems wrong, for at least two reasons: (i) patents play a different role in a completely public economy. The number of patents is hence an unsuitable benchmark, and (ii) the USSR did not only relevantly differ from the West with respect to their health care and drug development systems. Differences in drug development (and approval) are at least partly explained by a number of other differences between the systems, e.g. education, laws, policing, economy, incentive structures.
- Nancy Cartwright, together with a number of co-workers, has also been much interested in causal inference in the real world (e.g. [92,93]). Most of her studies focus on the social sciences. While these authors employ some medical examples, their approach does not constitute a fully fledged methodology for causal inference in medicine. Relevant to our purposes, Cartwright's work does highlight the issue of external validity for medical inference.
- Finally, there are a number of further approaches for evidence synthesis, which are in our assessment not fully-fledged accounts of causal inference with solid philosophical underpinnings such as [94–96] and see [97] for an overview.

3.1 EBM+

3.1.1 Approach

EBM+ is an alternative to EBM because it requires explicit evaluation of evidence of mechanisms alongside and on a par with evidence of correlation [11,98]. The motivation for EBM+ is an epistemological thesis made

8 See also [9,36,45,58,74–89].

by Russo and Williamson [99]: in order to establish a causal claim in medicine (A causes B), one normally needs to establish that A is appropriately correlated with B, and establish that a mechanism exists linking A and B that can account for the correlation. This thesis is sometimes called the Russo-Williamson thesis (RWT), and sometimes evidential pluralism.

According to EBM+ proponents, high-quality RCTs can establish causation because they provide indirect evidence that both a correlation and a mechanism exist, thus satisfying the conditions of RWT [100]. However, the problem EBM+ identifies with EBM_{epis} is that often there are enough issues with the implementation of RCTs that only a correlation is established. In such cases, establishing a mechanism can help, as the two lines of evidence reinforce one another: each line of evidence has a characteristic weakness that is addressed by the strengths of the other [101]. Most people outside the EBM+ community disagree with the EBM+ stance of treating evidence obtained from RCTs as mere correlational evidence. They instead uphold the epistemological importance of randomisation and the better quality evidence it provides.

EBM+ also places no restrictions on the kind of method used – causation can be established by evidence obtained from both clinical and mechanistic studies. The methods employed by mechanistic studies are various, but they are the methods of the "basic sciences" (e.g. microbiology, biochemistry, physiology, molecular biology). At best, in EBM_{epis}, evidence from mechanistic studies is taken as always strictly inferior to evidence obtained from clinical studies. But mechanistic studies can provide direct evidence of the existence of a mechanism [100]. According to EBM+, this is enough to establish a mechanism. So, a combination of clinical and mechanistic studies can suffice to establish causation. Recently, the consortium of philosophers and scientists at the heart of the EBM+ programme have produced a set of guidelines for assessing evidence from mechanistic studies [102]. This keeps in the tradition common to EBM_{epis} of systematically evaluating evidence, and operationalises the theory behind the approach.

3.1.2 Philosophical analysis

As noted earlier, there are many parties interested in whether a drug is approved. Reliable methods are supposed to be the way to counter this. But some authors argue that the methods of EBM_{epis} are not reliable and are in fact malleable – they can be bent to the will of researchers to produce the results they want – and this goes a long way to explain why the *status quo* is unsuccessful at drug approval (see, e.g., [103]). [104] argues that EBM+ as a tool for causal inference is actually less malleable than EBM_{epis} since (1) mechanistic studies and clinical studies reinforce each other as the strengths of each line of evidence makes up for the deficiencies of the other line of evidence (this is the basic argument for evidential pluralism); (2) mechanistic evidence is (to a large degree) independent from observational or intervention studies with respect to performing the studies, scientific instruments used in the studies, and the methodology used; and, (3) "Mechanistic studies ... are often conducted by research teams with different interests to those who carry out association studies (which tend to be carried out by drug companies seeking approval for lucrative new drugs). Thus publication bias, fraud, and industry manipulation are less of a concern for mechanistic studies than for association studies" [104]. For both benefit and harm assessment, we disagree. Instead, application of EBM+ to drug approval problems would plausibly lead to at least as bad outcomes as does EBM_{epis}.

Points 1 and 2 above are about the reliability of competing logics of causal inference – Williamson argues that EBM+ is more reliable than EBM_{epis}, hence will head off the problems EBM_{epis} faces. Now, one might think that the reliability claim is not true, e.g. it is not clear that independent evidence is more confirmatory. A growing literature on the variety of evidence thesis highlights that *less independent evidence is* (counter intuitively?) *more confirmatory* in a variety of cases: see [105–108] for the latest on such cases. However, even if we assume that (1) and (2) make the logic of the EBM+ causal evaluation more reliable than EBM_{epis}, it is not clear that (3) is true – the remainder of this analysis is aimed at rejecting the claim that "publication bias, fraud, and industry manipulation are less of a concern for mechanistic studies than for association studies."

Our argument starts by recognising that in an EBM+ world mechanistic studies *will be carried out* by drug companies seeking approval for lucrative new drugs. It is plausible that non-profit researchers have the monopoly on mechanistic studies only because there is currently no financial incentive to pharmaceutical

companies to be in the game. As theorised more abstractly in [70], if the rules (of the drug approval process) change, the behaviour of the stakeholders will likely change. At a basic level, more mechanistic studies means an increased chance of landing on what looks like a mechanism. This line of argument may falter when it comes to the search for harms, which already leans towards an EBM+ approach. The latest regulations for assessments of ADRs already explicitly call for the incorporation of mechanistic and other low-level evidence (Section 2.3.1). This implies that (i) strong enough evidence of a positive correlation between drug use and ADR or (ii) a great enough number of case reports can suffice for an assessment that the drug causes an ADR that would lead to drug disapproval. Drug companies would not want to increase the amount of mechanistic studies here. However, what they can do is exploit our often incomplete knowledge of biological mechanisms to argue that causation has not been established [109]. They can do this by producing conflicting evidence where the case is already strong that there is a plausible harm, or they can call into question the existing evidence where the case is weaker.

The proponent of EBM+ may argue in turn that strict enough assessment guidelines – such as Parkkinen et al. [102] – may solve this problem. In particular, that just having more evidence does not necessarily mean more mechanisms will be established. There must be evidence that the mechanism operates and this evidence must be good quality. EBM+ immediately runs into some problems. One is that companies will often manage to argue that there is enough evidence to establish a mechanism. If establishing a mechanism is a prerequisite for drug approval, then they will put their mind to establishing a mechanism. For one, they have the budget to carry out a great quantity of research, which is also much cheaper than conducting RCTs. Additionally, there are a great number of models for human (patho)-physiology on which to experiment: a great number of human or animal cell lines, different animal species, biomedical imaging technologies, autopsies, and computer simulations. The myriad of strict assessment frameworks available to regulators working in the *status quo* have not managed to ward off manipulation; why should EBM+ be any different? Furthermore, the multiplicity and heterogeneity of methods used to discover mechanisms is a weakness for EBM+ here rather than a strength. Approaches to mitigating industry bias may be difficult to implement for mechanistic studies. For example, pre-specifying a trial analysis is straightforward for an RCT, due to its relative simplicity. But it is very difficult to do this for any one mechanistic study, let alone all the different kinds of studies that can be carried out.

A final problem is the blind spot current EBM+ guidelines have for avoiding industry bias. Parkkinen et al. include no tools for assessing the extent of industry influence on the production of evidence. This point has recently been made by Howick [110] and is in contrast to the inclusion of such assessment criteria in some *status quo* frameworks, e.g. GRADE. Clearly, this gap in the guidelines is there because the EBM+ group think mechanistic studies are less likely to be subject to industry manipulation. As we have argued, there may be a greater chance of industry manipulation. This problem is exacerbated by the fact EBM+ delivers – by design – only qualitative judgements about the relative plausibility of the target causal relationship. One might worry that this leaves EBM+ open to the influence of subjective judgements. There is hence no reason to think that mechanistic studies will differ significantly from association studies with respect to bias, fraud and industry manipulation, and we conclude that implementing EBM+ would not succeed to any greater extent than the *status quo*.

Summing up, an implementation of EBM+ for drug approval faces plausible problems of applicability. EBM+ is, however, not the finished article, and we hope these critical points suggest new avenues for its proponents to explore.

3.2 E-synthesis

3.2.1 Approach

E-synthesis has been developed for safety assessments [111–117]. It aims to take into account all the available information and not just focus on RCTs, which is crucial for harm assessments in the real world (see Section 2.3.2). Typically, aggregated bodies of evidence are composed of diverse and contradictory safety signals.

In order to carry out this difficult task, design choices are made. α) E-Synthesis applies Bayesian epistemology which dominates modern analytic philosophy of science. β) E-Synthesis interprets the Bradford Hill Guidelines [118] as indicators of causation [119]. The basic idea is that an indicator being true increases our belief in the causal hypothesis, while an indicator being false decreases our belief. γ) E-Synthesis does not take information at face value: every item of evidence is assessed in terms of a number of evidential modulators, which modulate the extent of confirmation. For example, observational studies are assessed according to the number of observations, the duration of the study, the (strength of) sponsorship bias, the quality of adjustment for confounders and stratification, as well as external validity. δ) E-Synthesis then applies a Bayesian network approach [120] to calculate a posterior probability of a drug causing an ADR given the available and assessed evidence. In order to define the Bayesian network, the causal hypothesis, as well as every indicator of causation, every item of evidence and every modulator of every study[9] is represented by a variable. Strictly speaking, E-Synthesis is not wedded to any of these design choices, e.g. different indicators, modulators, and/or structures of the Bayesian network, could be used, yet implementing a number of such changes would result in a different approach worthy of a label different than E-Synthesis. We shall use the term E-Synthesis for the originally proposed approach.

E-Synthesis has not been developed for, nor has it ever been applied to, effectiveness assessments.

3.2.2 Philosophical analysis

E-Synthesis is at even earlier stage of development than EBM+: no comprehensive case study has been put forward, and no practical suggestions have been made for incorporating case reports into the evidence aggregation framework, although case reports play a major role in drug safety assessments in the real world [44]. We hence require a greater dose of imagination to envision how this approach would work in the real world.

Compared to the *status quo*, E-Synthesis holds the promise to systematically seamlessly aggregate evidence of different kinds: RCTs, observational studies, mechanistic evidence as well as case reports. This promise will, however, go unfulfilled, if the requirement to present the total evidence is not met (Section 2.4).

Furthermore, E-Synthesis makes a number of judgements explicit (e.g. assessed values of modulating variables, importance of indicators of causation) and allows their confirmatory roles to be tracked for hypothesis confirmation. This can be a significant improvement relative to the *status quo* in which safety assessments are – in the end – narrative reviews.

While the capability of explicity can be a major advantage, it also, we think, harbours a grave danger for sensible real world applications of E-Synthesis. Explicity is facilitated by the use of a great number of variables. The specification of the resulting Bayesian network required to calculate a posterior probability of a causal hypothesis of interest is hence carried out by the definition of a great number of probabilities. These probabilities are meant to represent an ideally rational agent's credences. Unfortunately, no ideally rational agent inhabits our world.

Setting these probabilities involves a good deal of subjective choice. While some think that subjectivity is not much of an issue [121], we are much more pessimistic. Our view is a consequence of the track record of rule exploitation, bending, and breaking, as well as the size of the incentives at stake (Footnote 3). To take some examples of subjective choice, (i) a prior probability of the causal hypothesis of interest must be specified; (ii) the strength of sponsorship bias of every item of evidence must be assigned a probability; (iii) the confirmatory value of all sets of indicators of causation must be formalised by defining conditional probabilities; (iv) all evidence modulating variables must be assessed; and (v) conditional probabilities of all items of evidence given the assessed modulating variables must be determined.

Any way of setting these probabilities in a scientifically defensible way will have to produce them in an objective sense, which would make use of all the other available information. For example, data from previous safety assessments could, in principle, be used to determine frequency information regarding the probabilities of indicators of causation, the strength of sponsorship bias, and the conditional probabilities of items of evidence given assessed sponsorship strengths. Distilling this information in the real world necessitates the

9 With the exception that some few evidence variables may share the same modulator variable.

solution of a greater number of choice problems, which do not have objectively correct solutions. For example, does one take into account studies concerning: the same/similar adverse event/reaction; same/similar study design; same/similar or even all drug manufacturers; same/similar drug design; same/similar group or researchers? These issues constitute a real world reference class nightmare [122]. Furthermore, there is a great number of ways to operationalise graded notions of the quality of adjustment, blinding, randomisation, and external validity.

We hence believe that every application of E-Synthesis to a drug approval problem would require a great number of choices to be made, where there are no obvious unique correct solutions. This entails that a great dose of subjectivity is required. In turn, this opens the floodgates for drug manufacturers to influence the outcome of safety assessments.

Overall, we believe that implementation of E-Synthesis would not lead to safer drugs but rather to more drugs being granted market approval and unsafer drugs as a consequence. Since the *status quo* is already under-assessing harms (Section 2.3), an implementation of E-Synthesis in its current form cannot be recommended.

3.3 Empowered patients

3.3.1 Approach

Before concluding, we now sketch a philosophical approach to drug approval that is neutral on how to draw causal inferences but differs on *who* ought to make decisions. Libertarian thinkers defend the view that patients experiencing the (good and bad) effects of drugs ought to be the ones who decide which drugs they take [123,124] (see [125] for more historical information on the paternalism vs freedom debate in drug approval).

3.3.2 Philosophical analysis

Teira recently argued that patients suffering ADRs will find it very hard to prove in court that they indeed suffered an ADR and hence are very unlikely to be compensated [126]. This argument is based on significant information asymmetries between plaintiffs and the defending corporation, as well as significant asymmetries to fund legal teams. We agree.

He went on to argue that drug regulation agencies ought to collect the available evidence and present it to the decision makers, the patients. We believe that this is unlikely to happen in the real world, since even the current regime fails to base their assessments on *all the available evidence* (Section 2.4).

Without access to a trusted impartial entity performing causal inference and decision making, real world patients will often fail to make good decisions, since they (1) typically lack the medical training to properly understand the medical details of studies; (2) are typically not trained in causal inference, and will thus struggle with properly assessing evidential support relations between causal hypotheses on the one side and observational and/or randomised studies on the other side; (3) are often ill and are thus not in a mental state to do difficult judgements, inferences and assessments; and (4) often do not have the time to read all the presented material.[10]

Overall, we strongly believe that the outcomes of all implementations of an empowered patients approach are significantly worse for patients than the *status quo*.

10 A very rough back-of-the-envelope calculation in previous study [127] estimates that a doctor would have to read “75 min of every working day, just for the medical education journals!” This calculation excludes reading primary evidence from academic journals. Granted, a patient will, at a time, only look at a small number of treatment options. However, (1) patients take, on average, a much longer time to read a medical article than a doctor due to not having trained for years in medical school; (2) The rate of publications of medical studies is ever increasing, the numbers from the study of Faux [127] are too small in today’s world. The number of citations on pubmed from 2000 - 2020 ca. 18,000,000, number of total citations 1800–2020 ca. 31,000,000. – these numbers were obtained by searching pubmed for, e.g. 1800:2100[dp].

4 Conclusions and recommendations

Our conclusions are based on an empirically informed methodology. Assessing the *status quo* systematically is infeasible in a philosophy study. Hence, we opted for an assessment methodology that tests the *status quo* against the evidence but, nevertheless, can draw no firm quantitative conclusions. Equally, assessing the alternative approaches based on their performance at drug approval is impossible. Instead, we carried out a philosophical analysis of application to drug approval and all its complexity.

In doing this, we followed the suggestion of [14] that drug approval decision-making procedures ought to be evaluated by how well their implementations perform. In order to do so, we specified assessment criteria (Section 2.1) and discussed how to operationalise them in principle (Section 2.5). Applying our assessment methodology, the *status quo* came out significantly worse than in the analysis of [14]. While the *status quo* was found to have serious problems in particular with respect to assessment criteria **A2** and **A3**, the alternative philosophical approaches were all found to be at least equally bad when contextualised in implementation to drug approval (Sections 3.1, 3.2). *As drug approval processes*, they can be viewed as somewhat worse than the *status quo*. While the logic of EBM+ and E-Synthesis' methods for friction free causal evaluation may have strengths over what the *status quo* offers, there are clear lacunae in each alternative with respect to drug approval. Furthermore, the empowered patients, approach is purely unsuitable in the same context. The CauseHealth, USSR, and Cartwright approaches (Section 3) cannot be evaluated within our assessment framework.

We would like to end with three reflections on the strengths and weaknesses of our analysis and then some recommendations for improving the *status quo*.

First, the two alternative approaches discussed in Sections 3.1 and 3.2 are both relatively novel. They arose from philosophical considerations and were at least initially not designed to be robust to clever manipulations by a multi-billion dollar industry. We think that further developments improving their robustness to manipulations is a promising avenue for further research. It thus seems premature to declare them out of contention for implementation. Second, drug approval is an important, but by no means the only important problem requiring causal medical inference. It will be interesting to assess how these approaches do when applied to other contexts. Third, although none of the philosophical approaches may do better than the *status quo* in drug approval decision procedures, we believe in the epistemological value of continuing to pursue well-grounded philosophical analyses. This is particularly so given continued calls for the aggregation of RWE that arises from outside RCT environments [17,128–130], possible applications to other concrete problems in medicine, and also for the purpose of yet to be discovered philosophical insights.

Our conclusions should not be read as saying the *status quo* should be jettisoned. It clearly does not fail outright and can be proud of some successes – a case in point is the recent testing and approval of very effective and safe COVID-19 vaccines. Instead, we believe recognising the issues and then addressing them is the best course of action for continuously evolving and improving drug approval procedures [18,131]. We certainly hope that the addendum from 2020 improves the *status quo*. Whether it does, is an empirical question we cannot yet answer.

In that vein, our recommendations are as follows:

R1 All drugs approved conditional on a follow-up study lose their approval, if the follow-up is not completed in good time or the results are not as positive as expected.
This rule, in principle, solves the problem of follow-up studies not being conducted (Section 2.4).

R2 All studies (including methodology and plan for data analysis) used in support of a drug must be pre-registered with the regulator, and all this evidence must be made available to the regulator prior to conducting the study.
This rule, in principle, solves the problem of violations of the Principle of Total Evidence and reporting biases by cherry picking evidence presented to the regulator.

R3 Set an expiry date for all approvals with a marginally positive risk-benefit balance.
This rule, in principle, solves the problem of drugs behaving different than anticipated (*e.g. less effective than* (*Section* 2.2.2) *and/or more dangerous* (*Section* 2.3.2)).

We write “in principle” because rules can be bent, exploited, and broken once implemented in the real world. Nevertheless, we hope that the suggested rules would go some way to address the problems they were designed to combat. Of course, this is not a complete list, and we welcome attempts to fashion further recommendations on the back of this article’s analysis and in line with the wider literature.

However, we can identify some possible costs to their implementation. Implementing R1 will mean that more follow-up studies are conducted, which cost effort and money and may put patients taking part in these studies in harm’s way. Setting expiry dates (R3) will also entail that more follow-up studies and more periodic reviews of drugs are conducted. We are echoing the call for more post-approval active controlled trials to better understand the comparative (dis-)advantages of newly approved drugs [25,54,132].

R2 is restricted to benefits since evidence of harms often emerges spontaneously in unpredicted and unpredictable ways. It is hence not sensible to circumscribe the search for and reporting of evidence of ADRs to pre-registered searches. Enforcing the Principle of Total Evidence for ADRs, too, would surely be beneficial for pharmacovigilance. We did not find an implementable rule that compels the disclosure of evidence that the regulator is not aware of. It is simply hard to compel people to carry out thorough searches for data from a priori unknowable locations/patients, collate all the so-obtained evidence, and use all of their brain power to squeeze out the last bit of information. Concerning the cost of implementation: pre-registration of studies and study methodology (R2) is already standard in many areas of medical research. In addition, dedicated protocols and IT infrastructure are already in place. So the use of R2 will cost only a tiny fraction of the millions of dollars it costs to develop a drug. While it might be necessary to deviate from a plan for data analysis [133], our proposal will at least require that all change of plans are made explicit and are argued for.

With that said, we strongly believe that incurring these (and other) costs is greatly outweighed by the benefits we all would accrue from the implementation of this study’s recommendations.

Acknowledgements: First and foremost, we owe a debt to Micheal Wilde, who was initially a co-author. Unfortunately, time constraints prevented him from seeing this article through to the end. We are grateful to Alexander Gebharter and Naftali Weinberger for valuable comments and suggestions.

Funding information: Jürgen Landes gratefully acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 528031869, 405961989 and 432308570 as well as NextGenerationEU funding for the project “Practical Reasoning for Human-Centred Artificial Intelligence.” We also gratefully acknowledge the Open Access Fund of the University of Milan covering the article processing charges.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: All data used are publicly available via the cited sources.

References

[1] La Caze A. Evidence-based medicine must be. J Med Philos. 2009;34(5):509–27. doi: https://doi.org/10.1093/jmp/jhp034.
[2] Howick JH. The philosophy of evidence-based medicine. Oxford, United Kingdom: Blackwell; 2011. doi: https://doi.org/10.1002/9781444342673.
[3] Worrall J. Evidence in medicine and evidence-based medicine. Philoso Compass. 2007;2(6):981–1022. doi: https://doi.org/10.1111/j.1747-9991.2007.00106.x.
[4] Stegenga J. Down with the hierarchies. Topoi. 2014;33(2):313–22. doi: https://doi.org/10.1007/s11245-013-9189-4.
[5] Osimani B. Hunting side effects and explaining them: should we reverse evidence hierarchies upside down? Topoi. 2014;33(2):295–312. doi: https://doi.org/10.1007/s11245-013-9194-7.
[6] La Caze A. Evidence-based medicine cant be... Soc Epistemol. 2008;22(4):353–70. doi: https://doi.org/10.1080/02691720802559438.

[7] Cartwright N, Munro E. The limitations of randomized controlled trials in predicting effectiveness. J Evaluat Clin Practice. 2010;16(2):260–6. doi: https://doi.org/10.1111/j.1365-2753.2010.01382.x.
[8] Solomon M. Just a paradigm: evidence-based medicine in epistemological context. Europ J Philos Sci. 2011;1(3):451–66. doi: https://doi.org/10.1007/s13194-011-0034-6.
[9] Holman B. Philosophers on drugs. Synthese. 2019;196:4363–90. doi: https://doi.org/10.1007/s11229-017-1642-2.
[10] Jones A, Steel D. Evaluating the quality of medical evidence in real-world contexts. J Evaluat Clin Practice. 2018;24(5):950–6. doi: https://doi.org/10.1111/jep.12983.
[11] Sung D, Holman B. Against evidential pluralism in pharmaceutical regulation. Philos Sci. 2023;90:1276–85. doi: https://doi.org/10.1017/psa.2023.40.
[12] Canadian Task Force. The periodic health examination. Canadian Med Assoc J. 1979;121(9):1193–254. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1704686/.
[13] Aronson JK, Caze AL, Kelly MP, Parkkinen VP, Williamson J. The use of mechanistic evidence in drug approval. J Evaluat Clin Practice. 2018;24(5):1166–76. doi: https://doi.org/10.1111/jep.12960.
[14] Andreoletti M, Teira D. Rules versus standards: what are the costs of epistemic norms in drug regulation? Sci Tech Human Values. 2019;44(6):1093–115. doi: https://doi.org/10.1177/0162243919828070.
[15] Luján JL, Todt O. Evidence based methodology: a naturalistic analysis of epistemic policies in regulatory science. Europ J Philos Sci. 2021;11(1):26. doi: https://doi.org/10.1007/s13194-020-00340-7.
[16] Van Norman GA. Drugs and devices. JACC Basic Translat Sci. 2016;1(5):399–412. doi: https://doi.org/10.1016/j.jacbts.2016.06.003.
[17] Sherman RE, Anderson SA, Pan GJD, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? New England J Med. 2016;375(23):2293–7. doi: https://doi.org/10.1056/nejmsb1609216.
[18] Committee for Medicinal Products for Human Use. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials; 2020. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf.
[19] Steel D. A new approach to argument by analogy: extrapolation and chain graphs. Philos Sci. 2010;77(5):1058–69. doi: https://doi.org/10.1086/656543.
[20] Bareinboim E, Pearl J. A general algorithm for deciding transportability of experimental results. J Causal Inference. 2013;1(1):107–34. doi: https://doi.org/10.1515/jci-2012-0004.
[21] Hernán MA, Vander Weele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011;22(3):368–77. doi: https://doi.org/10.1097/ede.0b013e3182109296.
[22] Pearl J, Bareinboim E. External validity: from do-calculus to transportability across populations. Stat Sci. 2014;29(4):579–95. doi: https://doi.org/10.1214/14-STS486.
[23] Kashoki M, Hanaizi Z, Yordanova S, Veselý R, Bouygues C, Llinares J, et al. A Comparison of EMA and FDA decisions for new drug marketing applications 2014-2016: concordance, discordance, and why. Clin Pharmacol Therapeut. 2020;107(1):195–202. doi: https://doi.org/10.1002/cpt.1565.
[24] Raphael MJ, Gyawali B, Booth CM. Real-world evidence and regulatory drug approval. Nat Rev Clin Oncol. 2020;17(5):271–2. doi: https://doi.org/10.1038/s41571-020-0345-7.
[25] Wieseler B, McGauran N, Kaiser T. New drugs: where did we go wrong and what can we do better? BMJ. 2019;366:l4340. Corrections available at https://doi.org/10.1136/bmj.l4837. doi: https://doi.org/10.1136/bmj.l4340.
[26] IQWiG. General Methods Version 6.0; 2020. https://www.iqwig.de/methoden/general-methods_version-5-0.pdf.
[27] Davis C, Naci H, Gurpinar E, Poplavska E, Pinto A, Aggarwal A. Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009–13. BMJ. 2017;359:j4530. doi: https://doi.org/10.1136/bmj.j4530.
[28] Kim C, Prasad V. Cancer drugs approved on the basis of a surrogate end point and subsequent overall survival. JAMA Int Med. 2015;175(12):1992. doi: https://doi.org/10.1001/jamainternmed.2015.5868.
[29] Gyawali B, Hey SP, Kesselheim AS. Assessment of the clinical benefit of cancer drugs receiving accelerated approval. JAMA Int Med. 2019;179(7):906. doi: https://doi.org/10.1001.
[30] van Luijn JCF, Gribnau FWJ, Leufkens HGM. Superior efficacy of new medicines? Europ J Clin Pharmacol. 2010;66(5):445–8. doi: https://doi.org/10.1007/s00228-010-0808-3.
[31] Kissin I. The development of new analgesics over the past 50 years: a lack of real breakthrough drugs. Anesthesia Analgesia. 2010;110(3):780–9. doi: https://doi.org/10.1213/ane.0b013e3181cde882.
[32] Mintzes B, Vitry A. Flawed evidence underpins approval of new cancer drugs. BMJ. 2019;366:l5399. doi: https://doi.org/10.1136/bmj.l5399.
[33] Kieffer CM, Miller AR, Chacko B, Robertson AS. FDA reported use of patient experience data in 2018 drug approvals. Therapeutic Innovat Regulat Sci. 2020;54(3):709–16. doi: https://doi.org/10.1007/s43441-019-00106-1.
[34] Van Norman GA. Update to drugs, devices, and the FDA. JACC: Basic Transl Sci. 2020;5(8):831–9. doi: https://doi.org/10.1016/j.jacbts.2020.06.010.
[35] Beasley CM, Dellva MA, Tamura RN, Morgenstern H, Glazer WM, Ferguson K, et al. Randomised double-blind comparison of the incidence of tardive dyskinesia in patients with schizophrenia during long-term treatment with olanzapine or haloperidol. British J Psychiatry. 1999;174(1):23–30. doi: https://doi.org/10.1192/bjp.174.1.23.

[36] Preston TA. DES and the elusive goal of drug safety. In: Dutton DB, editor. Worse than the disease: Pitfalls of medical progress. Cambridge: Cambridge University Press; 1988. p. 31–90.
[37] Food and Drug Administration. Drug induced liver injury: premarketing clinical evaluation - guidance for industry; 2009. http://www.fda.gov/downloads/Drugs/Guidance/UCM174090.pdf.
[38] Vandenbroucke JP, Psaty BM. Benefits and risks of drug treatments: How to combine the best evidence on benefits with the best data about adverse effects. JAMA. 2008;300(20):2417–9. doi: https://doi.org/10.1001/jama.2008.723.
[39] Vandenbroucke JP. When are observational studies as credible as randomised trials? The Lancet. 2004;363(9422):1728–31. doi: https://doi.org/10.1016/S0140-6736(04)16261-2.
[40] Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials. 2012;13(1):138. doi: https://doi.org/10.1186/1745-6215-13-138.
[41] Goldkind L, Laine L. A systematic review of NSAIDs withdrawn from the market due to hepatotoxicity: lessons learned from the bromfenac experience. Pharmacoepidemiol Drug Safety. 2006;15(4):213–20. doi: https://doi.org/10.1002/pds.1207.
[42] Duijnhoven RG, Straus SMJM, Raine JM, de Boer A, Hoes AW, Bruin MLD. Number of patients studied prior to approval of new medicines: a database analysis. PLoS Med. 2013;10(3):e1001407. doi: https://doi.org/10.1371/journal.pmed.1001407.
[43] Aronson JK. Post-marketing drug withdrawals: pharmacovigilance success, regulatory problems. Therapies. 2017;72(5):555–61. doi: https://doi.org/10.1016/j.therap.2017.02.005.
[44] Onakpoya IJ, Heneghan CJ, Aronson JK. Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis. Crit Rev Toxicol. 2016;46:477–89. doi: https://doi.org/10.3109/10408444.2016.1149452.
[45] Ehmann F, Papaluca-Amati M, Salmonson T, Posch M, Vamvakas S, Hemmings R, et al. Gatekeepers and enablers: how drug regulators respond to a challenging and changing environment by moving toward a proactive attitude. Clin Pharmacol Therapeutics. 2013;93(5):425–32. doi: https://doi.org/10.1038/clpt.2013.14.
[46] Brown JP, Wing K, Evans SJ, Bhaskaran K, Smeeth L, Douglas IJ. Use of real-world evidence in postmarketing medicines regulation in the European union: a systematic assessment of European medicines agency referrals 2013–2017. BMJ Open. 2019;9(10):e028133. doi: https://doi.org/10.1136/bmjopen-2018-028133.
[47] Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun JD, Krumholz HM, et al. Postmarket safety events among novel therapeutics approved by the US food and drug administration between 2001 and 2010. JAMA. 2017;317(18):1854. doi: https://doi.org/10.1001/jama.2017.5150.
[48] Salas-Vega S, Iliopoulos O, Mossialos E. Assessment of overall survival, quality of life, and safety benefits associated with new cancer medicines. JAMA Oncol. 2017;3(3):382. doi: https://doi.org/10.1001/jamaoncol.2016.4166.
[49] Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing regulation of medicines withdrawn from the market because of drug-attributed deaths: an analysis of justification. Drug Safety. 2017;40(5):431–41. doi: https://doi.org/10.1007/s40264-017-0515-4.
[50] Onakpoya IJ, Heneghan CJ, Aronson JK. Delays in the post-marketing withdrawal of drugs to which deaths have been attributed: a systematic investigation and analysis. BMC Med. 2015;13(1):26. doi: https://doi.org/10.1186/s12916-014-0262-7.
[51] Fornasier G, Francescon S, Leone R, Baldo P. An historical overview over pharmacovigilance. Int J Clin Pharmacy. 2018;40(4):744–7. doi: https://doi.org/10.1007/s11096-018-0657-1.
[52] Frank C, Himmelstein DU, Woolhandler S, Bor DH, Wolfe SM, Heymann O, et al. Era of faster FDA drug approval has also seen increased black-box warnings and market withdrawals. Health Affairs. 2014;33(8):1453–9. doi: https://doi.org/10.1377/hlthaff.2014.0122.
[53] Pease AM, Krumholz HM, Downing NS, Aminawung JA, Shah ND, Ross JS. Postapproval studies of drugs initially approved by the FDA on the basis of limited evidence: systematic review. BMJ. 2017;357:j1680. doi: https://doi.org/10.1136/bmj.j1680.
[54] Naci H, Salcher-Konrad M, Kesselheim AS, Wieseler B, Rochaix L, Redberg RF, et al. Generating comparative evidence on new drugs and devices before approval. The Lancet. 2020;395(10228):986–97. doi: https://doi.org/10.1016/S0140-6736(19)33178-2.
[55] Salcher-Konrad M, Naci H, Davis C. Approval of cancer drugs with uncertain therapeutic value: a comparison of regulatory decisions in Europe and the United States. Milbank Quarterly. 2020;98(4):1219–56. https://onlinelibrary.wiley.com/doi/10.1111/1468-0009.12476.
[56] Arku D, Yousef C, Abraham I. Changing paradigms in detecting rare adverse drug reactions: from disproportionality analysis, old and new, to machine learning. Expert Opinion Drug Safety. 2022;21:1–4. doi: https://doi.org/10.1080/14740338.2022.2131770.
[57] Jefferson T, Jones M, Doshi P, Spencer EA, Onakpoya I, Heneghan CJ. Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. BMJ. 2014 Apr;348(2):g2545–5. doi: https://doi.org/10.1136/bmj.g2545.
[58] Christian A. On the suppression of medical evidence. J General Philos Sci. 2017;48(3):395–418. doi: https://doi.org/10.1007/s10838-017-9377-9.
[59] Dyer O. Cochrane reviewer sues Roche for claiming Tamiflu could slow flu pandemic. BMJ. 2020;368:m314. doi: https://doi.org/10.1136/bmj.m314.
[60] Carnap R. On the application of inductive logic. Philos Phenomenol Res. 1947;8(1):133–48. doi: https://doi.org/10.2307/2102920.
[61] Seruga B, Templeton AJ, Badillo FEV, Ocana A, Amir E, Tannock IF. Personalising drug safety-results from the multi-centre prospective observational study on adverse drug reactions in emergency departments (ADRED). Europ J Clin Pharmacol. 2016;76(3):439–48. doi: https://doi.org/10.1007/s00228-019-02797-9.
[62] Bavli I, Steel D. Inductive Risk and OxyContin: the ethics of evidence and post-market surveillance of pharmaceuticals in Canada. Public Health Ethics. 2020;13(3):300–13. doi: https://doi.org/10.1093/phe/phaa031.

[63] Zorzela L, Golder S, Liu Y, Pilkington K, Hartling L, Joffe A, et al. Quality of reporting in systematic reviews of adverse events: systematic review. BMJ. 2014;348(jan08 1):f7668–8. doi: https://doi.org/10.1136/bmj.f7668.
[64] Demasi M. FDA oversight of clinical trials is "grossly inadequate," say experts. BMJ. 2022;379:o2628. doi: https://doi.org/10.1136/bmj.o2628.
[65] Isakov L, Lo AW, Montazerhodjat V. Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design. J Econometrics. 2019;211(1):117–36. doi: https://doi.org/10.1016/j.jeconom.2018.12.009.
[66] Intriligator MD. Drug evaluations: type I vs type II errors; https://escholarship.org/uc/item/5fg9n284.
[67] Eichler HG, Bloechl-Daum B, Brasseur D, Breckenridge A, Leufkens H, Raine J, et al. The risks of risk aversion in drug regulation. Nature Rev Drug Discovery. 2013;12(12):907–16. doi: https://doi.org/10.1038/nrd4129.
[68] Mueller S, Pearl J. Personalized decision making - A conceptual introduction. J Causal Inference. 2023;11(1):20220050. doi: https://doi.org/10.1515/jci-2022-0050.
[69] Wardell WM. Introduction of new therapeutic drugs in the United States and great Britain: an international comparison. Clin Pharmacol Therapeutics. 1973;14(5):773–90. doi: https://doi.org/10.1002/cpt1973145773.
[70] Holman BH. The fundamental antagonism: science and commerce in medical epistemology; 2015. PhD Thesis at UC Irvine. https://escholarship.org/uc/item/4kx8g2r1#author.
[71] Herresthal C. Hidden testing and selective disclosure of evidence. J Econ Theory. 2022;200:105402. doi: https://doi.org/10.1016/j.jet.2021.105402.
[72] Henry E, Ottaviani M. Research and the approval process: the organization of persuasion [CEPR Discussion Papers]. Am Econ Rev. 2019;109(11939):911–55. doi: https://doi.org/10.1257/aer.20171919.
[73] Ross JS, Hill KP, Egilman DS, Krumholz HM. Guest authorship and ghostwriting in publications related to rofecoxib: A case study of industry documents from rofecoxib litigation. J Am Med Assoc. 2008;299(15):1800–12. doi: https://doi.org/10.1001/jama.299.15.1800.
[74] John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 2012;23(5):524–32. doi: https://doi.org/10.1177/0956797611430953.
[75] Horton R. Vioxx, the implosion of Merck, and aftershocks at the FDA. The Lancet. 2004;364(9450):1995–6. doi: https://doi.org/10.1016/S0140-6736(04)17523-5.
[76] Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. The Lancet. 2004;364(9450):2021–9. doi: https://doi.org/10.1016/S0140-6736(04)17514-4.
[77] Nestle M. Corporate funding of food and nutrition research: Science or marketing? JAMA Int Med. 2016;176(1):13–4. doi: https://doi.org/10.1001/jamainternmed.2015.6667.
[78] Holman B, Geislar S. Sex drugs and corporate ventriloquism: how to evaluate science policies intended to manage industry-funded bias. Philos Sci. 2018;85(5):869–81. doi: https://doi.org/10.1086/699713.
[79] Ioannidis JPA. Evidence-based medicine has been hijacked: a report to David Sackett. J Clin Epidemiol. 2016;73:82–6. doi: https://doi.org/10.1016/j.jclinepi.2016.02.012.
[80] González-Moreno M, Saborido C, Teira D. Disease-mongering through clinical trials. Stud History Philos Sci Part C Stud History Philos Biol Biomed Sci. 2015;51:11–8. doi: https://doi.org/10.1016/j.shpsc.2015.02.007.
[81] Holman B, Bruner J. Experimentation by industrial selection. Philos Sci. 2017;84(5):1008–19. doi: https://doi.org/10.1086/694037.
[82] Holman B, Elliott KC. The promise and perils of industry-funded science. Philos Compass. 2018;13(11):e12544. doi: https://doi.org/10.1111/phc3.12544.
[83] Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Library. 2017;2:MR000033. doi: https://doi.org/10.1002/14651858.MR000033.pub3.
[84] Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. JAMA. 1998;279(19):1566–70. doi: https://doi.org/10.1001/jama.279.19.1566.
[85] Bekelman JE, Li Y, Gross CP. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. JAMA. 2003;289(4):454–65. doi: https://doi.org/10.1001/jama.289.4.454.
[86] Bes-Rastrollo M, Schulze MB, Ruiz-Canela M, Martinez-Gonzalez MA. Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: a systematic review of systematic reviews. PLOS Med. 2013;10(12):1–9. doi: https://doi.org/10.1371/journal.pmed.1001578.
[87] Ioannidis JPA. Hijacked evidence-based medicine: stay the course and throw the pirates overboard. J Clin Epidemiol. 2017;84:11–3. doi: https://doi.org/10.1016/j.jclinepi.2017.02.001.
[88] Pham-Kanter G. Revisiting financial conflicts of interest in FDA advisory committees. Milbank Quarterly. 2014;92(3):446–70. doi: https://doi.org/10.1111/1468-0009.12073.
[89] Sismondo S. Ghost management: how much of the medical literature is shaped behind the scenes by the pharmaceutical industry? PLoS Med. 2007;4(9):e286. doi: https://doi.org/10.1371/journal.pmed.0040286.
[90] Anjum RL, Copeland S, Rocca E, editors. Rethinking causality, complexity and evidence for the unique patient. Cham: Springer; 2020. doi: https://doi.org/10.1007/978-3-030-41239-5.
[91] Rocca E, Anjum RL. Causal evidence and dispositions in medicine and public health. Int J Environ Res Public Health. 2020;17(6):1813. doi: https://doi.org/10.3390/ijerph17061813.
[92] Cartwright N, Hardie J. Evidence-based policy. Oxford: Oxford University Press; 2012.

[93] Deaton A, Cartwright N. Reflections on randomized control trials. Soc Sci Med. 2018;210:86–90. doi: https://doi.org/10.1016/j.socscimed.2018.04.046.
[94] Dammann O. Evidence mapping to justify health interventions. Perspectives Biol Med. 2021;64(2):155–72. doi: https://doi.org/10.1353/pbm.2021.0018.
[95] Verde PE. A bias-corrected meta-analysis model for combining, studies of different types and quality. Biometric J. 2020;63(2):406–22. doi: https://doi.org/10.1002/bimj.201900376.
[96] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–64. doi: https://doi.org/10.1093/aje/kwv254.
[97] Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. Res Synthesis Methods. 2014;6(1):45–62. doi: https://doi.org/10.1002/jrsm.1122.
[98] Greenhalgh T, Fisman D, Cane DJ, Oliver M, Macintyre CR. Adapt or die: how the pandemic made the shift from EBM to EBM+ more urgent. BMJ Evidence-Based Med. 2022;27(5):253–60. doi: https://doi.org/10.1136/bmjebm-2022-111952.
[99] Russo F, Williamson J. Interpreting causality in the health sciences. Int Stud Philoso Sci. 2007;21(2):157–70. doi: https://doi.org/10.1080/02698590701498084.
[100] Williamson J. Establishing causal claims in medicine. Int Stud Philoso Sci. 2019;32(1):33–61. doi: https://doi.org/10.1080/02698595.2019.1630927.
[101] Auker-Howlett D, Wilde M. Reinforced reasoning in medicine. J Evaluat Clin Practice. 2020;26(2):458–64. https://onlinelibrary.wiley.com/doi/abs/10.1111/jep.13269.
[102] Parkkinen VP, Wallmann C, Wilde M, Clarke B, Illari P, Kelly MP, et al. Evaluating evidence of mechanisms in medicine: principles and procedures. Cham, Switzerland: Springer; 2018. doi: https://doi.org/10.1007/978-3-319-94610-8.
[103] Stegenga J. Medical Nihilism. Oxford: Oxford University Press; 2018.
[104] Williamson J. The feasibility and malleability of EBM+. THEORIA. 2020;36(2):191–209. doi: https://doi.org/10.1387/theoria.21244.
[105] Osimani B, Landes J. Varieties of error and varieties of evidence in scientific inference. British J Philos Sci. 2023;74(1):117–70. doi: https://doi.org/10.1086/714803.
[106] Landes J. The variety of evidence thesis and its independence of degrees of independence. Synthese. 2021;198:10611–41. doi: https://doi.org/10.1007/s11229-020-02738-5.
[107] Landes J. Variety of evidence and the elimination of hypotheses. Europ J Philos Sci. 2020;10:12. doi: https://doi.org/10.1007/s13194-019-0272-6.
[108] Casini L, Landes J. Confirmation by robustness analysis. A bayesian account. Erkenntnis. 2024;89:367–409. doi: https://doi.org/10.1007/s10670-022-00537-7.
[109] Plutynski A. Explaining cancer: finding order in disorder. New York: Oxford University Press; 2018. http://www.oxfordscholarship.com/view/10.1093/oso/9780199967452.001.0001/oso-9780199967452.
[110] Howick J. Exploring the asymmetrical relationship between the power of finance bias and evidence. Perspectives Biol Med. 2019;62(1):159–87. doi: https://doi.org/10.1353/pbm.2019.0009.
[111] Landes J, Osimani B, Poellinger R. Epistemology of causal inference in pharmacology. Europ J Philos Sci. 2018;8:3–49. doi: https://doi.org/10.1007/s13194-017-0169-1.
[112] De Pretis F, Landes J, Peden WJ. Artificial intelligence methods for a bayesian epistemology-powered evidence evaluation. J Evaluat Clin Practice. 2021;27(3):504–12. doi: https://doi.org/10.1111/jep.13542.
[113] De Pretis F, Peden WJ, Landes J, Osimani B. Pharmacovigilance as personalized evidence. In: Bertolaso M, Canali S, editors. Personalized medicine in the making. Cham: Springer; 2022. p. 147–71. doi: https://doi.org/10.1007/978-3-030-74804-3_8.
[114] De Pretis F, Landes J, Osimani B. E-Synthesis: a Bayesian framework for causal assessment in pharmacosurveillance. Front Pharmacol. 2019;10:1317. doi: https://doi.org/10.3389/fphar.2019.01317.
[115] Abdin Y, Auker-Howlett DJ, Landes J, Mulla G, Jacob C, Osimani B. Reviewing the mechanistic evidence assessors e-synthesis and EBM.: a case study of amoxicillin and drug reaction with Eosinophilia and systemic symptoms (DRESS). Curr Pharm Des. 2019;25(16):1866–80. doi: https://doi.org/10.2174/1381612825666190628160603.
[116] De Pretis F, Osimani B. New insights in computational methods for pharmacovigilance: e-synthesis, a Bayesian framework for causal assessment. Int J Environ Res Public Health. 2019;16(12):2221. doi: https://doi.org/10.3390/ijerph16122221.
[117] De Pretis F, Landes J. A softmax algorithm for evidence appraisal aggregation. PLoS ONE. 2021;16(6):1–23. doi: https://doi.org/10.1371/journal.pone.0253057.
[118] Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58(5):295–300.
[119] Bovens L, Hartmann S. Bayesian epistemology. Oxford: Oxford University Press; 2003.
[120] Neapolitan RE. Learning Bayesian networks. Upper Saddle River: Pearson; 2003.
[121] Sprenger J. The objectivity of subjective Bayesianism. Europ J Philos Sci. 2018;8:539–58. doi: https://doi.org/10.1007/s13194-018-0200-1.
[122] Hájek A. The reference class problem is your problem too. Synthese. 2007;156:563–85. doi: https://doi.org/10.1007/s11229-006-9138-5.
[123] Reiss J. Meanwhile, why not biomedical capitalism? In: Elliott KC, Steel D, editors. Current Controversies in Values and Science. New York: Routledge; 2017. p. 161–75. doi: https://doi.org/10.4324/9781315639420.
[124] Flanigan J. Pharmaceutical freedom: why patients have a right to self medicate. Oxford: Oxford University Press; 2017.

[125] Fraile Navarro D, Tempini N, Teira D. The trade-off between impartiality and freedom in the 21st Century Cures Act. Philos Med. 2021;2(1). doi: https://doi.org/10.5195/philmed.2021.24.
[126] Teira D. A defence of pharmaceutical paternalism. J Appl Philos. 2020;37:528–42. doi: https://doi.org/10.1111/japp.12413.
[127] Faux D. Information overload. Medical Teacher. 2000;22(1):5–6. doi: https://doi.org/10.1080/01421590078724.
[128] ECETOC. Framework for the integration of human and animal data in chemical risk assessment; 2009. http://www.ecetoc.org/uploads/Publications/documents/TR.
[129] Rocca E, Copeland S, Edwards IR. Pharmacovigilance as scientific discovery: an argument for trans-disciplinarity. Drug Safety. 2019;42(10):1115–24. doi: https://doi.org/10.1007/s40264-019-00826-1.
[130] Review of EPA's integrated risk information system (IRIS) process. Washington: National Academies Press; 2014. doi: https://doi.org/10.17226/18764.
[131] European Commission. Proposal for a regulation amending, as regards pharmacovigilance of medicinal products for human use. Regulation (EC) No 726/2004; 2008. http://ec.europa.eu/health/files/pharmacos/pharmpack_12_2008/pharmacovigilance-ia-vol1_en.pdf.
[132] Cipriani A, Ioannidis JPA, Rothwell PM, Glasziou P, Li T, Hernandez AF, et al. Generating comparative evidence on new drugs and devices after approval. The Lancet. 2020;395(10228):998–1010. doi: https://doi.org/10.1016/S0140-6736(19)33177-0.
[133] Dutilh G, Sarafoglou A, Wagenmakers EJ. Flexible yet fair: blinding analyses in experimental psychology. Synthese. 2021;198:5745–72. doi: https://doi.org/10.1007/s11229-019-02456-7.