Predictive validity in drug discovery: what it is, why it matters and how to improve it

Jack W. Scannell[1], James Bosley[2], John A. Hickman[3], Gerrard R. Dawson[4], Hubert Truebel[5], Guilherme S. Ferreira[6], Duncan Richards[7] and J. Mark Treherne[8]

Affiliations

1. Science, Technology, and Innovation Studies, University of Edinburgh, J.W. Scannell Analytics LTD, Unify Pharmaceuticals Corp. jack.scannell@btinternet.com
2. NovaDiscovery SA, jim.bosley@novadiscovery.com
3. School of Biological Sciences, Manchester University, jhickman@hotmail.fr
4. P1vital Ltd, gdawson@p1vital.com
5. The Knowledge House GmbH, hubert.truebel@knowledge-house.com
6. 3D-PharmXchange, gsantferr@gmail.com
7. Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford. duncan.richards@ndorms.ox.ac.uk
8. Talisman Therapeutics Ltd, mark@talisman-therapeutics.com

Corresponding author: J.W. Scannell, +44 7576661611 or jack.scannell@btinternet.com

Abstract | Successful drug discovery is like finding oases of safety and efficacy in chemical and biological deserts. Screens in disease models, and other decision tools used in drug research and development (R&D), point towards oases when they score therapeutic candidates in a way that correlates with clinical utility in humans. Otherwise, they probably lead in the wrong direction. This line of thought can be quantified by using decision theory, in which 'predictive validity' is the correlation coefficient between the output of a decision tool and clinical utility across therapeutic candidates. Analyses based on this approach reveal that the detectability of good candidates is extremely sensitive to predictive validity, because the deserts are big and oases small. Both history and decision theory suggest that predictive validity is under-managed in drug R&D, not least because it is so hard to measure before projects succeed or fail later in the process. This article explains the influence of predictive validity on R&D productivity and discusses methods to evaluate and improve it, with the aim of supporting the application of more effective decision tools and catalysing investment in their creation.

[H1] Introduction

Drug discovery involves decisions. We decide the likelihood that the development of a particular therapeutic candidate — for example, a putative drug target or chemical compound — will lead to a useful human drug. Some candidates then receive further investment and others are abandoned. The decisions often apply to a large number of candidates at successive stages of the research and development (R&D) process. Decisions often involve huge uncertainty. The challenges are immense. Indeed, a lack of sufficiently predictive methods for target validation, and for identifying and optimising therapeutic candidates, is now seen as the main technical bottleneck in drug discovery[1–5].

In this paper, we regard "decision tools" (Table 1) as all the things that are used to score therapeutic candidates in a way that is believed to correlate with clinical utility, to help decide to optimize and advance some candidates, and to abandon others. Decision tools include target-based and phenotypic screens, mechanistic physiological simulations, artificial intelligence (AI), tests in disease models, the 'gut feel' of experienced scientists and experimental medicine methods in human participants, as well as the documents and processes used by management teams to support their decisions to progress with drug R&D projects.

In this context, a decision tool's 'predictive validity' (Table 1) is the degree to which the scores it gives to a set of therapeutic candidates (for example, putative targets or drugs) correlates with clinical utility in people. With this formalism, it has been shown previously by two of the authors (Scannell and Bosley) that changes in predictive validity that appear trivially small and that are rarely practical to measure — for example, shifting the correlation coefficient from 0.5 to 0.6 — are surprisingly important in quantitative terms[6]. Such changes could often have a bigger impact on the probability of selecting a clinically useful drug candidate by using the decision tool (its positive predictive value (PPV); Table 1) than a 10× or sometimes a 100× change in the number of candidates tested[6]. If one takes an aggregate view and treats all the activity leading up to a phase I go/no-go decision as a "meta-decision tool", the same small improvements in aggregate predictive validity could be worth tens, perhaps hundreds, of millions of dollars per phase I candidate (see later).

Thus, the thing that nearly everyone already believes is important is more important than nearly everyone already believes.

The aim of this Perspective is to set out what might be done practically to nudge predictive validity in the right direction; both for drug discovery and preclinical development technologies and for the management processes that aggregate evidence from them. It starts by summarising the case for predictive validity's importance by looking at historical trends in R&D productivity and cases where

changes in decision tool validity transformed the R&D enterprise. Next, we overview the decision-theoretic logic presented in the previous in-depth analysis[6] and its implications (including the financial implications for an archetypal[7] big pharma pipeline), with the aim of making these accessible to a broad audience. We then consider the kind of education and training that could help organizations take a more validity-centric view, how to prospectively evaluate predictive validity, the implications for R&D management and investment, and finally, the problematic economics of decision tool-related innovation.

We largely avoid the pros and cons of specific technologies, except when illustrating our validity-related thinking. We are not going to speculate on the general prospects for artificial intelligence (AI)[8,9], biobanking and deep phenotyping[10], patient or volunteer-derived models[5], nor a host of other innovations. This is not because we think these technologies do not matter, nor because we think the pharma industry is locked into old-fashioned approaches. Our focus is for two reasons. First, we think that the R&D productivity debate is often too technology-centric, when really it should start from position that is decision-centric and technology-agnostic. Second, the predictive validity of R&D technologies is domain-specific. This makes it hard to say much that is sensible about a range of technologies in an article of reasonable length. However, we do provide some decision-theoretic speculation on phenotypic versus target-based drug discovery, and on the role of serendipity in pharmaceutical innovation in the supplementary information.

[H1] Importance of decision tool validity

The decline in pharmaceutical R&D productivity that occurred between around 1950 and 2010 is well known and has received considerable attention[3,7,19,20,11–18]. But it is still striking, in our view, when set against huge gains in scientific knowledge and the hundred, thousand, or even billion-fold improvements in the brute-force efficiency of many of the technologies that are believed to be important. Given such contrasting trends in input and output efficiency, the major causative factors presumably acted progressively over time and should account for orders of magnitude of efficiency change[12].

Scannell and Bosley[6] argued that the quantitative power of predictive validity is such that it could be a major causative factor. They hypothesized that the best assays and disease models identified good drugs, which succeeded in clinical trials, entered widespread clinical use, became generic, raised the commercial barriers to new R&D, and so rendered the best models industrially redundant. Think, for example, of models of stomach acid secretion[21,22] and the safe and effective H_2 receptor antagonists and proton pump inhibitors that they identified. Poor assays and models, on the other hand, failed to identify enough good drugs so, ironically, remained in widespread use, sometimes for decades. Think,

for example, of cancer-derived cell lines[23–25], durably high attrition rates[22,26,27], and the modest clinical benefit of most approved cytotoxic[28–32] and oncogene-targeted cancer treatments[33,34]. The gradual decline in the predictive validity of the stock of models in widespread use could offset the brute force efficiency gains. Support for this line of thinking is evident in a paper by Shih et al. (2017)[22], which looked at attrition rates across 15,000 therapeutic-mechanism–indication pairs, drawn from the Cortellis pipeline database. They found, for example, very high mechanism–indication pair success rates for proton pump inhibitors and very low success rates in oncology, suggesting that "the lack of translatability from strong therapeutic evidence in animal models into human disease therapeutics [points to] the need to carefully evaluate the predictive power of such animal models."

In parallel, we suspect that much of the pharmaceutical industry sometimes made the wrong technological trade-offs because it had not understood the quantitative power of predictive validity. It sometimes embraced discovery methods with measurably high throughput and low unit costs, whose benefits were offset by less measurable falls in predictive validity. A clear example is antibacterial R&D (Box 1). *In vivo* phenotypic screens of a few hundred compounds, circa 1930[35–37], were more productive than target-based screens of ~10^7 compounds in the late 1990s and early 2000s[35,37–41].

If a fall in predictive validity can explain strong R&D headwinds, then marginal gains could provide a meaningful tailwind. Indeed, the creation of good decision tools has transformed therapy in some domains. R&D on direct-acting hepatitis C antiviral drugs was impractical until the invention of HCV "replicons"[42–44] (see Box 2) that made it possible to produce reliable, high-titre, viral RNA replication in cell cultures, and to screen and optimize drugs[45]. Recent industry-wide trends are also suggestive. Ringel et al.[46] argued that the uptick in FDA drug approvals since 2010 follows, in part, from genetic segmentation that matches targets, translational models and patient populations, a focus on genetically simple rare diseases that can be modelled with relatively high fidelity, and a stronger managerial focus on "truth-oriented" versus "progression-oriented" activities[1,3,47,48].

[H1] Decision theory applied to drug R&D

[H2] A common-sense model of drug R&D

Decision theory has a long history in drug R&D, but the foci have usually been technical problems, often related to clinical trial design[49–57] at one end of the process, or chemistry and screening at the other[58,59,68,60–67] (but see, for example, reference[69]).

Scannell and Bosley sought to make a broader application — to the R&D process as a whole[6]. They were aware that the efficiency of search methods can be very sensitive to the details of the search task and it seemed reasonable to think of drug R&D as a kind of search. This approach led to a

mathematical representation of R&D to explore the parameters that influence search efficiency, formalising a common-sense view of R&D[6] (Figure 1). It assumes: (1) that there are, or could be created, a great many therapeutic candidates. These could be the compounds that might be screened and the derivatives that could be synthesized during lead optimization, that antibodies that might be created, or the universe of drug targets that might be relevant for a particular disease; (2) that randomly selected therapeutic candidates are unlikely to have enough clinical utility (based on properties such as safety, efficacy and convenience) to work in people. For practical purposes, "unlikely" means less than around 10%, so the results are more relevant for discovery and preclinical development and less relevant from phase I trials onwards in drug development; (3) that there are various methods for estimating the clinical utility of the candidates, before it is measured definitively in human trials, which we call "decision tools" (Table 1); (4) that decision tools vary in the degree to which their outputs correlate with clinical utility in people (that is, their predictive validity) and with the outputs of other decision tools; and (5) that therapeutic candidates' performance on the decision tools is set against performance thresholds, implicit or explicit, to decide which candidates progress. For a longer discussion on the assumptions, and their relevance to real drug R&D, see the supplementary information.

We operationalize predictive validity as the Pearson correlation coefficient (Table 1) between the output of the decision tool and clinical utility in people across a wide range of therapeutic candidates, but it could be operationalized in other ways (for example, Spearman rank correlation coefficient or area under an ROC curve; see Table 1). Predictive validity is an aggregate quantitative measure of the performance of a decision tool, so it subsumes factors such as reproducibility and replicability. One should also think about predictive validity in decision-making terms, not purely technical terms. A management committee's go/no-go deliberations, or a venture capital firm's investment process, count as "decision tools" because they estimate the utility of a set of opportunities, to invest in some and not others. If decision makers call for the wrong information, are biased[70], or are careless, then their predictive validity can fall far below the limit set by the information to which they have access.[71]

[H2] Key results from theory application

Scannell and Bosley explored the decision performance of a single R&D step involving a single decision tool. They also looked at more realistic situations by combining tools and steps in series and in parallel. They varied the degree to which decision tool outputs correlated with each other as well as with clinical utility. They explored a range of different distributions of clinical utility and decision tool scores. Across a wide range of conditions there was a consistent message: quality beats quantity. When good candidates are rare, changes in decision tools' predictive validity that many working scientists would regard as small and/or unknowable (for example, an absolute 0.1 change in the

correlation coefficient between decision tool output and clinical utility in people) can drive larger changes in a decision step's PPV (Table 1, Figure 2) than testing 10× or even 100× as many candidates (Figure 3). This follows from fewer false positives when predictive validity is higher, but also from more true positives and fewer false negatives[6] (compare the false negative and true positive quadrants in Figure 3a versus Figure 3b).

If one treats all discovery and preclinical activity as an aggregate decision tool (see the supplementary information, where we justify this approach), then small changes in predictive validity can be worth tens or even hundreds of millions of dollars of expected net present value (NPV; Table 1) per drug candidate entering phase I (Figure 3); this is similar to the effect of a two-fold change in the profit from each drug that R&D ultimately delivers (Figure 3). It is worth emphasizing this point. Statistics text books and primers[72–74] lump together correlation coefficients; for example, 0.7 to 0.9 is "high," 0.5 to 0.7 is "medium", and 0.3 to 0.5 is "low". This implies that there is not much practical difference within each group. In our experience, some biologists share this view. A 0.7 correlation coefficient does not really "feel" so different to a 0.5, not least because sample sizes running into the hundreds would be needed to confidently distinguish between the two[75]. However, the R&D productivity difference resulting from the use of an 'aggregate decision tool' with a predictive validity for clinical utility of 0.7 rather than 0.5 could be worth hundreds of millions of dollars per phase I candidate (Figure 3).

[H1] Practical implications

Of course, people working in drug R&D already know that decision tool validity is extremely important[22]. Some industrial strategies have converged on a "validity-centric" approach[3,76]. There is obvious enthusiasm for human genetic validation of potential drug targets[2,77–80]. There is a plethora of work on, for example, translational strategies within companies[1,76] , assay development[81], experimental reproducibility, reporting quality and bias[82–91], and even on getting rid of the worst of the old disease models[92]. There is also some evidence that this has started to increase the number of drug approvals[46].

However, both the decision theoretic work and the real-world examples (Boxes 1 to 3, and the oncology and mood disorder examples discussed later) suggest there will often be value in more rigorous and explicit consideration of the predictive validity of the decision tools being used during drug R&D programmes. In what follows, we consider several aspects of the challenge of achieving this goal —education and awareness, decision tool evaluation, and approaches to R&D management and investment — before concluding with our thoughts on how to fund the creation of better decision tools.

[H2] Education and awareness

The training of most biological scientists is scant on validity-related and decision-theoretic ideas. We provide a reading list in the supplementary information. In brief, however, we recommend that, first, there should be an understanding of the quantitative importance of decision tool validity[6,86] and what this means for effective R&D.

Second, within organizations, there should be a *lingua franca* for important validity-related ideas. Current terminology is inconsistent and confusing (for an illustration, see Belzung and Lemoine[93] for five different definitions of "construct validity" from the psychopharmacological literature). Converging on a *lingua franca* should force some clarity and make it easier to manage the R&D process.

Third, people should understand that conventional hypothesis testing and p-values are problematic when applied in drug discovery and preclinical development[94,95]. They can be prolific false discovery generators[94]. Statistical concepts that apply to diagnostic testing[96,97] are often more appropriate in our view. Biomedical scientists should perhaps be as familiar with ROC curves[96], confusion matrices[96], true positive rates, false positive rates, positive predictive values and false discovery rates (Table 1) as they are with p-values and hypothesis tests.

And fourth, people should be trained in decision tool evaluation, to improve their ability to identify, create, and exploit results from better decision tools and to discount results from worse ones[98,99].

[H2] Decision tool evaluation

Those who have read up to this point may have started to think that predictive validity is to drug R&D as "dark matter" is to cosmology[100]. It is important. It has a strong theoretical basis. It explains a range of empirical observations (for example, those discussed in Boxes 1, 2 and 3), but it largely defeats direct measurement. If so, how can one evaluate it? And without evaluation, how can one manage it or invest on its basis?

Perhaps a useful analogy here is that of the difference between Bayesian and frequentist approaches to probability. Bayesian probability deals with degrees of belief about what will happen. Frequentist probability deals with historical counts of events. Choices about screens used in drug R&D usually have a Bayesian flavour because they are based on beliefs — often implicit and untested[99] — that the outputs correlate with clinical utility. Without such beliefs, then one might as well make R&D decisions via the toss of a coin or the roll of a dice.

We are not arguing for an impractically "frequentist" approach to the measurement of the predictive validity. The universe of therapeutic candidates is large. Only a tiny fraction is tested in people and that fraction contains only true positives and/or false positives from the discovery and preclinical stages. The false negatives and true negatives never enter the clinic. Therefore, one can rarely measure the correlation coefficient between human clinical utility and the output of any decision tool across a representative sample of therapeutic candidates.

We are arguing for evaluation, or measurement, to move the "Bayesian" view of decision tools a bit closer to the underlying reality. Recall that measurement concerns the reduction, not elimination, of uncertainty[101–104]. Decision tool evaluation makes practical sense if the improvement in decision performance from lower uncertainty is worth more than the cost of measurement[101]. **Error! Reference source not found.**~~Figure 4~~ shows that small reductions in uncertainty regarding predictive validity can be worth a great deal. We may also need less measurement than we think. When uncertainty is high, it can be materially reduced with little information[101]. Uncertainty around decision tool validity *is* often high, and the effort expended on evaluation has often been low[99]. There are probably correlates of decision tools' predictive validity that are observable[77,78,110–112,79,93,98,105–109]. There is a readable management literature[101,104], reflecting a large technical literature[102,103,113–119], on how to tackle difficult evaluation tasks, and on how to decide if the benefit of evaluation is likely to outweigh the costs[101] (Figure 4). But this knowledge has been under-deployed, perhaps because the value of measurement is not clear unless one runs the decision theoretic maths[6,86] (Figure 3, Figure 4).

Of course, what is known about validity, the value of additional information and the practicality of obtaining it, varies enormously, so the measurement approach should vary enormously too. However, it is possible to make some evidence-based generalisations: evaluation tends to be more effective when information is structured[101], consistently presented, attempts are made to calibrate the measurement (whether subjective or objective)[104,117,120], the exercise is iterative[121], and care is taken to involve the right mix of experts[113,116] and to reduce bias[122].

[H3] Evaluation structure. There are often gains from breaking down the main question (that is, "how valid is this decision tool?") into a hierarchical set of sub-questions. This structure makes it easier to provide the background information that is required. It makes it harder for evaluators to forget features that might be important[123]. It also makes it easier to explain to others why the decision tool is believed to be good, bad or indifferent[102]. Disaggregation can force a degree of common understanding, or overt disagreement, on the features of the decision tool that are important. It can also highlight differences between decision tools that would otherwise appear similar. For these

reasons, disaggregation is a common approach to subjective or quasi-subjective evaluation, with evidence to support its effectiveness in certain situations, particularly when uncertainty is high[101,118,123–126].

There is good starting material. The reproducibility crisis[127,128] and failed translation[85,129–132] have spawned a literature that disaggregates good practice with respect to experimental conduct and reporting[87–91,132–136]. There are breakdowns to assess the utility of models of disease[93,98,105–108,111] and conceptually similar breakdowns on the translatability of projects into the clinic[106,110,137,138] and the quality of drug targets[4].

We have been influenced by the "Framework to Identify Models of Disease" (FIMD)[98], which was developed by one of the authors of this paper. FIMD is oriented towards animal models, but we think the principles are applicable to other kinds of model. FIMD assesses the fit between the model and the human clinical state across eight domains: epidemiology, natural history and symptomatology, genetics, biochemistry, aetiology, histology, response to known pharmacology (positive and negative), and experimental endpoints. FIMD also separately considers reporting quality and the risk of experimental bias.

To give a sense of the effort required to apply FIMD, in the case of Alzheimer's disease it would take an established Alzheimer's model expert several weeks to tailor the evaluation criteria to the specific features of the human disease state, and then about a week or two of work per rodent model evaluated, depending on data availability.

For ease of communication, we condensed the FIMD domains under three general categories (Table 2). The first category is biological recapitulation, or the extent to which the decision tool captures relevant aspects of the biology of the clinical state. We are not arguing that more complex models are necessarily better. A simplified system can have excellent performance if it recapitulates the relevant biology (Box 2). However, it is important to know which aspects are captured and which are not. The second category is tests and endpoints, or the extent to which testing and scoring the therapeutic candidates is relevant to the way candidates will be tested and scored in clinical trials. This also considers the score the candidates need to achieve to be "hits" or "yesses" and proceed to the next stage of the R&D process. The third category is statistical and experimental hygiene, which considers factors such as experimental conduct, reporting quality, managed heterogeneity[139–141], bias, reproducibility and statistical power.

There is then a fourth category, often missing from evaluation frameworks, which is qualitatively different, and which aims to assess and explain the kind of decisions that a decision tool can

reasonably support: the decision tool's 'domains of validity'[6,108], or the parameters within which decision tool output is likely to be predictive. The idea that models have specific domains of validity is common in science. FIMD, for example, considers "context of use"[98]. However, this is too rarely made explicit in drug R&D in our view. In physics, for example, classical mechanics has excellent predictive validity for the motion of objects that are not hugely massive, that are not approaching the speed of light and that are bigger than atoms. When things are too massive and too fast, general relativity is a better model. When things are too small, quantum mechanics has higher predictive validity. Explicit consideration of domains of validity helps one understand a decision tool's strengths and weaknesses. It can also help one assemble a combination of models to plug evidential gaps.

There will be further breakdowns under each of these headings, tailored to the specific evaluation task in hand. Table 2 is oriented towards animal models. However, we illustrate the framework with both *in vitro* and animal-based models (Boxes 1–3 and later). With high-throughput screening early in the discovery process, there might be fewer and different criteria for tests and endpoints. Statistical and experimental hygiene, while important, should be relatively unambiguous. The main cognitive effort would focus on the sufficiency of biological recapitulation and the extent to which the decision tool's domains of validity include clinical utility in people (see Box 1). Historical studies using animal models of ischaemic stroke, which led to a slew of failed development programmes, would often have benefitted from more scrutiny of tests and endpoints, among other things (see Box 3). A recent study of the translatability of pharmacokinetic and pharmacodynamic results from animals to humans[142] points to challenges around tests and endpoints. Failed translation from mouse studies in amyotrophic lateral sclerosis (ALS) highlighted problems with statistical and experimental hygiene[85].

[H3] Calibrating subjective judgements. Disaggregation yields, in effect, a checklist. Deciding the relative importance of the items on the checklist, and how a decision tool scores on each item, will usually have a large subjective component. Fortunately, there is a large body of experience on subjective evaluation[101–104,113,115].

For example, training can help evaluators understand the ratings scales and their own biases[122]. There is a variety of methods, which could be delivered via case studies on established decision tools whose performance is regarded as good (for example, the spontaneously hypertensive rat[143] for antihypertensive drugs), bad (for example, mouse amyloidosis as a model of moderate to advanced Alzheimer's disease[144]), or indifferent (for example, rodent swim tests for antidepressants[112]).

Where practical, evaluators should be given "base rate" information[117,120,145]; information that bears on the historical performance of similar decision tools in similar contexts. It may be important to know,

for example, that projects supported by human genetic validation are roughly twice as likely to succeed as those that are not[79], but that not all genetic validation is equal[78]. Mendelian genetic evidence appears much more predictive than evidence from genome-wide association studies (GWAS)[78]. And Mendelian evidence is much more predictive when the trait closely resembles the putative drug indication[78].

Quantitative point estimates plus confidence intervals are generally better than ordinal ratings even if the person doing the rating is highly uncertain about their quantitative estimates[101,103]. A response of the type "I believe that the scores from this decision tool would correlate 0.7 with the human clinical outcome across the drug candidates of interest, with a correlation of 0.4 as the lower boundary of my subjective 90% confidence interval, and a correlation of 0.8 as the upper boundary." is much more useful than one of the type "I believe that the score from the decision tool is a '4' on a scale that runs from '1' (useless) to '5' (excellent).".

The ordinal '4' can mean a wider range of things to different people than can the 0.7 correlation coefficient. The 0.4 to 0.8 confidence intervals summarise subjective uncertainty. The ordinal '4' does not. It is easier to sense-check quantitative estimates which, in turn, can improve evaluation. After all, some assays, when repeated, do not correlate with themselves at the 0.7 level, let alone with human clinical utility. Where one is stuck with ordinal scales, ranks should be well defined and illustrated with real-world examples.

[H3] Reducing bias: evaluate against a target profile. The drug industry has huge experience in minimizing bias in some of its activities. The primary endpoints of pivotal trials are pre-specified as is the performance standard that is required to declare success. Treatment allocation and evaluation are often blinded.

Were one to apply similar principles to important decision tool evaluations, one would assemble a disinterested group of experts and run a decomposition exercise[106] to produce a tailored version of FIMD or Table 2. Against each sub-question, one would elicit a working definition of good decision tool performance. Think of this as the decision tool equivalent of a target product profile. Any given decision tool could then be compared against the profile, preferably by the same experts, before deciding how much to invest in believing its results. Throughout the process, one would take care to minimize the well-documented biases[117,122] that reduce the quality of subjective evaluation.

People who routinely scrutinize opportunities in a particular therapy area might want to use this kind of process to derive a set of target decision tool profiles (for example, venture capital firms and drug companies' business development teams). However, in our experience, things that resemble target

profiles are rare. One example comes from an industry-led consortium that is developing standards for microphysiological systems[146] (such as organs-on-a-chip and organoids). Their work on drug-induced liver injury (DILI) approaches a target profile[105] against which *in vitro* DILI models can be evaluated. It specifies target urea and albumin production per million hepatocytes, gene expression profiles, several biomarkers of normal liver function, certain histological features, and performance against a set of drugs that are either positive controls (that is, known to cause DILI in people) or matched negative controls (drugs from the same classes that are less toxic). An evaluation against these specifications would, in our terminology, mainly be an evaluation of biological recapitulation (Table ~~Table~~ 2). But even this impressive example[105] lacks detail that is required to put it into practice, particularly with respect to tests and endpoints (Table 2). It has little guidance on dosing the positive and negative controls. Nor does it specify the methods one should use to decide the extent to which output is congruent with the test compounds' known toxicity. Even "known toxicity" is a slippery concept, with a range of operational definitions[147].

Of course, it will usually be impractical to assemble groups of experts to do the decomposition, pre-specification and decision tool scoring. R&D project teams will generally be the ones who set standards and do the evaluation. But even here, there are ways of improving performance through training and calibration, through the design of incentives, by including decision tool evaluation in project progression decisions or by occasional third-party audits of teams' evaluation activity.

[H3] A scorecard or a score? Across a range of domains, from NASA's project risk assessment to forecasting student's exam grades, numerical decision algorithms generally beat human experts[101,104,148]. However, when we are trying to evaluate decision tools, "scorecards" will often be more useful than single validity scores, produced by aggregating across the different evaluation criteria (Table 2).

This is for at least three reasons. First, the pattern of a decision tools' strengths and weaknesses is probably itself informative[98], particularly when thinking about its domains of validity or combining information. Second, the data will not support easy aggregation. Evaluations are likely to use ordinal ratings (for example, from "1" = "useless" to "5" = "excellent"). An ordinal "4" is not the sum of two ordinal "2s". This makes numerical aggregation difficult, and bad aggregation can be worse than no aggregation[101]. Third, a lot of data are required to train and test numerical algorithms. Even with a mere 3 numerical scores (for example, for biological recapitulation, tests and endpoints, and statistical and experimental hygiene), one would need to evaluate in the order of 70 different decision tools, and have an independent objective measure of how good or bad each one was, to produce a regression-based decision model (see the sample size calculator in Related links).

[H3] Tackling the difficult problem of feedback. Evaluation performance is more likely to improve if there is a feedback loop so that those who are judging the models can find out whether they were right or wrong and understand why[104,121]. The drug R&D process presents unusual problems here. First, there is a great deal of project attrition that has nothing to do with the decision tools in question. Second, it takes years, sometimes decades, to find out if a decision tool was sufficiently predictive. Third, the incentives to perform adequate post-mortems may be too weak.

We have three suggestions to help. The first suggestion is to fund retrospective studies to understand why some decision tools give us the right answer and why others get it wrong. For example, use one set of archival data to design an evaluation process and then a matched but previously unseen set of archival data to test evaluation performance. Can one distinguish features of the decision tools that were more likely to lead correct decisions? A historical approach has been used to show the positive influence of human genetic validation on likelihood of success in clinical development[78,79], to help design AstraZeneca's 5Rs framework for improving R&D productivity[3] and to look at the origin of first-in-class drugs with a view to understanding the relative contributions of different discovery strategies[149,150].

A starting point could be the paper noted earlier by Shih et al.[22], which assessed R&D failure and success rates across pairs of therapeutic mechanisms and therapeutic indications: ~10,000 drugs representing ~2,400 therapeutic mechanisms tested in ~1,400 indications. In some mechanism–indication pairs, drug candidates had 100% success rates (for example, proton pump inhibitors for duodenal ulcers, gonadotropin-releasing hormone agonists for breast cancer and VEGF inhibitors for renal cell carcinoma), while in others, the failure rate was 100% (for example, NMDA receptor antagonists for cerebrovascular disease and ACAT inhibitors for atherosclerosis). What was different about the decision tools[22]? We are also struck that some failures of translation see public post-mortems so that lessons can be learned (for example, ischaemic stroke in Box 3, antimicrobials in Box 1), while other major campaigns fail spectacularly but quietly fade away (for example, IGF1 receptor blockers in oncology[151,152]).

The second suggestion is forward looking and concerns institutional learning. This is an argument for starting decision tool evaluation soon, for revisiting the historic evaluations as projects succeed or fail, and then using what is learned to train scientists, improve decision tools, and to inform technology choices and project progression decisions. We understand that this kind of learning led to Lipinski's "Rule of Five" for small-molecule oral bioavailability[62].

The third suggestion is to invest to structure, capture and share detailed protocols, calibration methods, calibration results, decision tools' contexts of use and associated "craft skills"[102] across organizations. There are some pre-competitive and some commercial offerings that are reminiscent of what we are suggesting here (see the Collaborative Adverse Outcome Pathway, DataFAIRy Bioassay Annotation Wiki and Springer Nature Experiments in Related links).

[H3] Oncology and decision tool evaluation. In this section we illustrate some evaluation ideas with two examples from oncology: cytotoxic drugs, which dominated the R&D effort from the 1940s until the 1990s, and oncogene-targeted drugs, which became widespread from around 2000. We do not consider immuno-oncology because it has only become a major field recently and it also lies further from the direct experience of the authors. We note, however, that the successes that led to immuno-oncology's current prominence in cancer R&D owe little to the mainstream oncology discovery technologies of the time.

The first effective systemic therapy for cancer followed the observation that exposure to mustard gas caused leukopenia in people[153–155]. This led, in 1942, to the first encouraging treatment of a patient with leukaemia. Chemical refinement created a portfolio of alkylating agents[153] (for example, chlorambucil, melphalan and cyclophosphamide). The refinement of the alkylating agents, and subsequent cytotoxic drugs, depended on fast-growing cell lines that were adapted to grow in culture or as grafts in rodents. The US National Cancer Institute (NCI) routinely screened compounds in two murine leukaemia cell lines from 1955[23,25], shifting to 60 human-derived cancer cell lines in 1985[24].

How well did the decision tools perform? Well, cytotoxic cancer drugs had among the highest clinical trial failure rates of any major therapy area[27,155] and the drugs that emerged gave little benefit to the majority of patients with advanced cancer[28–32]: "it is clear that cytotoxic chemotherapy only makes a minor contribution to cancer survival"[29]. Yet there were some triumphs such as curative treatments in rare cancers[30]; for example, childhood leukaemias, Hodgkin's lymphoma and testicular cancer.

A shift in the cancer research paradigm began around 1976[156] with the emergence of the oncogene concept[157,158], but oncogenes remained a minority interest in the drug industry until the early 2000s. Then, the landmark approvals of the HER2-targeted antibody trastuzumab (Herceptin) for HER2-positive breast cancer in 1998 and the BCR–ABL kinase inhibitor imatinib (Gleevec) for chronic myelogenous leukaemia (CML) in 2001 (both barely emerging from battles for institutional support[159,160]) catalysed a major change in the focus of anticancer R&D. Cancer was now often viewed through the lens of "oncogene addiction"[161,162], with the products of oncogenes amenable to targeted therapy.

However, the cancer models in academic and industrial use shifted less than the paradigm. Rapidly growing cell lines in culture and in xenografts remained important, later supplemented with genetically engineered mouse models[163] (GEMMS) with switchable on–off expression of oncogenes, and with xenografts of human tumour tissue grafted into immune-deprived mice[164,165] (PDX models). Outcome measures continued to focus on short-term (for example, 35-day) tumour growth[166,167].

How well did the assays perform? Between 2000 and 2015, oncology had the highest clinical trial failure rate of any major therapy area[26]. The drugs that emerged from trials failed to yield the predicted[168] therapeutic revolution for cancer in general[33,34,169–175]. As of 2018, around 5% of patients[34] diagnosed with advanced cancer in the US could expect to benefit from an FDA-approved oncogene-targeted therapy, and among those patients, resistance to treatment generally evolves[176]. But again there were some real triumphs (such as imatinib[177,178]) and a range of major advances in subsets of cancers (such as ALK[179] and EGFR[180] inhibitors in subsets of patients with lung cancer).

How might more formal decision tool evaluation have helped? At least it could have made it more obvious where R&D would fail. Turning to cytotoxic drugs, work to develop a target decision tool profile would have recognized that cancers vary hugely in their doubling times and cellular growth fractions[181–183]. Some haematological malignancies double in mass every 1 to 4 days. Some colon cancers double in 10 weeks or more. That this predicts response to cytotoxic therapy has been clear since the 1960s[182,183]. Growth kinetics also affect the therapeutic index of cytotoxic drugs, which tend to be more poisonous to normal tissues with high growth fractions than to cancers with low growth fractions. Cancers also vary in their resistance to programmed cell death (apoptosis) following the DNA damage that cytotoxic drugs cause; a fact whose clinical relevance was spotted in the 1970s[184], but which was largely ignored until the 1990s[185,186].

Decision tools were generally short experiments in fast-growing, high-growth-fraction, cell lines that did not recapitulate the biology of slow growing or apoptosis-resistant cancers. It should be no surprise, therefore, that the drugs the tools identified worked well in the rare cancers dominated by fast-growing, high-growth-fraction, populations of cells with low apoptosis resistance (for example, childhood leukaemias) and had only limited use in most other cancers. The models recapitulated the biology of a small subset of cancers and their domains of validity extended little further than the small subset of cancers whose biology they recapitulated.

Turning to oncogene-targeted therapies, it has been known since the 1970s[187] that very few cancers are monoclonal, where one event drives uncontrolled proliferation. The majority are polyclonal, with different subclones carrying different oncogenic drivers. We have learned that genetic instability

makes cancers less treatable[188]; about their complex evolution[189,190]; about multiple drivers of malignancy[191,192]; and topographic genetic heterogeneity within single cases of advanced cancer[191].

These features are missing in most oncology models. Cell lines are genetically homogeneous in two-dimensional growth or when used to construct three-dimensional models such as spheroids and tumour organoids, and none of these models capture the self-sculpted complexity of, and reciprocity between, a tumour and its microenvironment[193]. GEMMs have lacked the genetic heterogeneity necessary to evolve resistance to drug treatment[187,191], although recent models strive to capture the complex genomic damage characterising an advanced tumour[194]. In fact, when, in 2010, a GEMM *was* engineered to have more realistic genetic heterogeneity, via engineered chromosomal instability, the anticancer benefit of suppressing expression of its classical oncogenic driver mutation largely vanished[195]. Human tumour-derived xenografts appear more realistic in several important respects[111,165], but still suffer constraints[111,164], not least their tiny volume. Small early-stage, surgically resectable tumours in people have a volume of several cubic centimetres[196]; about one thousand times larger than the 2–3 cubic millimetres[197] of patient-derived tissue that would be grafted into a mouse. This means individual grafts are unlikely to capture the genetic heterogeneity[191,192] that tends to make advanced cancer resistant to treatment. Of course, it is possible to make multiple grafts into multiple mice, but this in turn raises other problems; not least 'propagation bias' as many — often the overwhelming majority — of human tumour fragments fail to grow[111,164] in their new host. There is also bias at the level of the individual graft, as minor clones from the original tumour often come to populate the model[198,199].

As with the cytotoxic drugs, the prevalent decision tools found good oncogene-targeted therapies for the relatively small proportion of cancers whose biology the decision tools recapitulated; cancers with dominant driver mutations and relatively stable and homogeneous genetics. But as with the cytotoxic drugs, most advanced cancers fall outside the decision tools' domains of validity, so relatively few patients benefitted from the oncogene-targeted drugs the decision tools selected[34]. This is illustrated by imatinib in CML. As long as CML resembles a monoclonal hyperplasia, not an advanced cancer[200], imatinib is typically effective. Once CML evolves into a typically genetically heterogeneous cancer (for example, during blast crisis), imatinib becomes ineffective[200].

Could decision tool evaluation have done more than help avoid failed oncology trials? We see two possible opportunities. The first is that such an evaluation could have provided support for a simpler approach to cancer R&D. The EMA[201] and FDA[202] permit phase I cancer trials with relatively little preclinical efficacy data, with drug candidate selection based on "the target and mechanisms of action," plus safety pharmacology, pharmacokinetics and toxicology to support a human dosing plan[202]. In our experience, some companies are reluctant to take a light-touch approach, preferring to

invest heavily in poorly predictive animal efficacy data. This may be because review boards and clinicians expect such data before recruiting patients into trials, even though the efficacy data they see often carries little useful information[142,203]. However, when there is no opportunity to substantially improve decision performance, minimizing cost is the best policy[69].

The second opportunity could have been to focus decision tool development at a different point on the trade-off between unit cost and predictive validity (analogous to Box 1, see also Figure 3) by investing more to produce tumour models that capture native growth kinetics, complexity and genetic heterogeneity. There has been some progress with human tumour slices[204,205] and "cancers on a chip"[206], but there are technical challenges. Tumour slices in static culture, without vasculature, have limited durability. Chips mimic vasculature and some aspects of tissue structure, but cancer chips have generally been produced with genetically homogenous cancer-derived cells[206].

[H2] R&D management and investment

Predictive validity should carry due weight in decision-making on project progression and investment[98,101,102,106,110]. Estimates of the attrition rate of subsequent R&D steps should be flexed up or down depending on the validity of the preceding decision tools (see the financial analyses in the supplementary materials). Management should encourage evaluation-based decision tool choice and discourage choice via tradition and availability bias[99,207]. Decision tool quality will often beat quantity[6] (Figure 3). There should be particular focus on early steps in the R&D process, since improvements in predictive validity when good drug candidates are rare have a bigger impact on productivity than improvements of a similar magnitude later in the process when good candidates are more common. Decision processes should be designed to minimize the loss of information between those who have practical 'craft knowledge' of the decision tools and those who make major spending and progression decisions.

Several of the authors have seen major funding decisions that depended on decision tool results where validity-related information was not merely insufficient, it was entirely absent from the documents provided to the decision-making group. A recent review of 28 Investigator Brochures for compounds entering phase I trials found that less than one third contained any text that sought to justify the choice of animal efficacy models[142]. None discussed the animal efficacy model choice versus other options that would have been available. Wieschowski et al. observed similar results[203].

Given the financial sensitivities (Figure 3, **Error! Reference source not found.**~~Figure 4~~), drug companies' audit and risk management committees should scrutinise the decision tool evaluation process when it bears on major acquisitions. Institutional investors (for example, pension funds and sovereign wealth

funds) should understand the decision tool evaluation methods of the biotech venture capital firms that invest on their behalf. Perhaps institutional review boards and clinical investigators should also make similar demands before enrolling patients into phase I and phase II trials[142,203].

[H3] Pharmacological calibration and cross-calibration. For important decision tools, the likely costs and benefits[101] of pharmacological calibration should be estimated (Figure 4) and then, where it is cost-effective, a set of compounds should be put through the tools.

The most obvious case is to measure predictive validity by testing the decision tools with a sample of drugs whose known clinical utility varies widely. We know of examples using toxicology models[108,147,208] and cancer cell lines[209–212], and it should be practical, if often expensive, in therapy areas where a large number of drugs have gone into human trials. This provides a quantitative measure of predictive validity and helps to set decision thresholds; the score on the model that best corresponds to a go or no-go decision on a candidate. Models with acceptable predictive validity may give the wrong decisions because of untested, yet easily testable, assumptions about the optimal go/no-go threshold.

As the practice of counter-screening[213,214] shows, one can also make better decisions if one knows whether the results of different decision tools are highly correlated, and thus redundant, or orthogonal, and thus potentially synergistic with each other[6,215]. Therefore, there can be value in cross-calibrating different decision tools using the same set of compounds[108,147]. To give an example, the lack of congruence between certain cancer cell line-based models has been presented in the academic literature as a problem[209–212], with critics arguing that congruence is low, and supporters arguing that congruence is high. In fact, pooled results from the different models would be more useful for decision making[6] if the critics are correct.

[H3] Plausibly higher predictive validity as a trigger for R&D investment. Human genetic validation of a potential drug target is an obvious example of something that raises predictive validity[2,77–80]. It has become a trigger, and for some firms a prerequisite, for R&D investment. However, it is hard to find genetically validated targets that are not obvious to other people. This encourages competitive crowding in oncology[216] and rare diseases, offsetting the financial benefits to the firms involved[46,217].

It should be possible to avoid the crowds by looking for correlates of predictive validity where competitive pressure is lower; perhaps in human tissue collections[5], micro-physiological systems[146,206,218], mechanistic simulation[219–223] and comparative physiology[224,225]. The projects for which the model systems promise the most reliable prediction could then be funded.

We illustrate this line of thinking with the Fast-MAS programme in mood and anxiety disorders[226–228], a response to the withdrawal of much of the drug industry from psychiatric R&D, which is funded by the US National Institute of Mental Health. It is an example of a larger body of work that is aiming to improve predictive validity in psychiatry via better pathological taxonomy[229–231] (which makes it easier to map screening and disease models onto psychiatric symptoms) and via new experimental medicine methods (which eliminate the species gap between some model systems and the patient).

From a commercial perspective, mood disorders have become a difficult place to allocate R&D capital. Many widely used antidepressants are available as cheap generic versions, which reduces demand for new and more expensive drugs. R&D is difficult. Human genetics has not yet helped much[232]. Preclinical models are often poor[112], hence the critical role of serendipitous clinical observations (akin to "phenotypic screening" in people) — which have low throughput but high validity — in the discovery of the major antidepressant classes[17,233] (see the supplementary information for further discussion of serendipity). Conventional clinical trial economics are further undermined by high placebo response rates (which reduce the signal to noise ratio), by high drop-out rates, by treatment responses that can take up to 8 weeks to emerge, plus a range of practical problems around patient selection[234,235]. In our terminology, even pivotal trials in depression have struggled with statistical and experimental hygiene[234].

Fast-MAS began by defining the criteria for a high validity development programme and only then searched for therapeutic candidates, targets and compounds to which the criteria could be applied. The idea was that by defining better models — in this case, experimental medicine methods — targets and compounds become investible[227].

Fast-MAS is tackling anhedonia, or the inability to enjoy normally pleasurable activities. Anhedonia is a common component of mood and anxiety disorders. Perhaps the first validity-related innovation here was the choice of clinical problem. Conventional psychiatric diagnosis is problematic, based on collections of symptoms and not necessarily aligned with underlying pathobiology[229,230]. This decorrelates disease models, which often assume a particular mechanism, from clinical utility within a conventional diagnostic group[229,230]. Anhedonia, on the other hand, appears to be something of a coherent pathobiological entity[226,228].

Tools to show target engagement were also a prerequisite for target and drug choice. In this case, there was a PET tracer available for the κ-opioid receptor (KOR), for which activation had been implicated in anhedonia in animals. The PET tracer had already revealed good target engagement in humans by a KOR antagonist, JNJ-67953964; a drug candidate that also appeared to be acceptably safe for clinical use.

Methods to show "proof of mechanism" were the next prerequisite; something with a high signal-to-noise ratio, hence measurable in a small phase IIa trial, and directly linked to the therapeutic mechanism. In this case, several strands of evidence from the neurobiology of reward and punishment led to an fMRI-based approach focused on part of the brain known as the ventral striatum. Around 90 subjects with anhedonia were randomized to receive JNJ-67953964 or placebo and were, near the start and end of treatment, placed in an fMRI scanner. Once there, they were given three alternative cues before being required to press a button with, depending on the cue, the button press either earning them money, avoiding a loss, or having no consequence[226]. The fMRI outcome measure depended on the difference in striatal blood flow between the states of financial anticipation, fear, or indifference. JNJ-67953964 normalized anhedonic subjects' fMRI measures versus placebo[226].

The validity-related innovations in Fast-MAS do seem to have triggered new investment by others[227]. Takeda adopted a similar fMRI approach to anhedonia in patients with schizophrenia. Janssen put JNJ-67953964 into a more conventional phase II trial as an adjunctive treatment in major depressive disorder (MDD).

There are, of course, some challenges. By drug industry standards, the meticulous Fast-MAS project was anything but fast; partly owing to the statistical and experimental hygiene challenges for a new fMRI-based endpoint collected across many different trial sites. Furthermore, anhedonia, while perhaps reflecting underlying biological reality, does not map neatly onto the diagnoses with which the FDA is comfortable. This means that companies must either take a regulatory risk on a new indication, or else run trials in conventional diagnostic categories, such as MDD or schizophrenia, where treatment may have only a weak effect on conventional rating scales.

Nonetheless, better pathological taxonomy plus behavioural biomarkers may make psychiatric R&D more tractable, as genetic taxonomy and molecular biomarkers have done in other therapy areas. The broader point is that human genetics is merely one route to the higher predictive validity that tends to raise returns on R&D investment.

[H1] A wider economic context

We have argued that the sensitivity of R&D productivity to small changes in the predictive validity of decision tools is underappreciated. We have discussed things that might nudge predictive validity in the right direction, via education, decision tool evaluation, and by giving predictive validity an overt role in project progression and investment. But decision tool-related innovation suffers from a general

economic problem: weak incentives for investment when the inventor cannot capture enough of the wider value that the invention creates[236,237].

At least for the diseases of the rich, the pharmaceutical R&D knowledge bottleneck has arguably shifted over time from chemistry to biology[1–5]. Some of the most predictive decision tools may be commercially exhausted[6]. Chemical space, on the other hand, is hardly scratched by the last 100 years of drug discovery[238,239]. Yet it is easier to capture financial value from novel drug structures than from novel decision tools. The value of compounds can be captured via composition-of-matter patents, even with relatively little[240] evidence that the compounds will be clinically useful. In contrast, if one invests in better decision tools to prove the worth of a new target, or to identify the tiny proportion of compounds that have clinical utility against that target, much of the value of the innovation can spill over to other firms at low cost; for example, when results from early human trials become public or when staff move between firms. This weakens incentives for decision tool-related innovation versus compound-related innovation.

Our view is based on several strands of evidence. It is suggested by economic analysis by De Villemeur and Versaeval[241]. It is exemplified in hepatitis C, where the profitability of novel chemistry — Gilead's oral antiviral drug sofosbuvir (Sovaldi) — caused political controversy[242,243]. Few politicians have heard of the replicons that made the novel chemistry possible (Box 2). It is shown by investors' beliefs about future profits, which drive company valuations. As of March 2022, the market capitalisation of the top 10 global contract research organisations was around 7.5% of the market capitalisation of the top 10 global pharmaceutical companies; investors expect companies that produce or acquire novel chemical matter to be able to capture far larger future profits than companies that supply the services that help decide whether the chemical matter is useful or not. It is apparent in biotechnology-related capital flows and in the compound-oriented innovation that those capital flows support[244,245].

This is a difficult problem to fix. The first step, of course, is to characterize the problem in more detail. We think that changing intellectual property laws would be ferociously difficult and, perhaps, not very effective. There is an argument for a shift in public sector R&D funding. However, academic science is not generally set up to perform the expensive, often mundane, long-term work needed to build, calibrate and maintain decision tools.

One class of solution is R&D collaboration between drug companies to fund work to improve the predictive validity of the suite of decision tools. The authors of this paper have some experience of public–private partnerships and precompetitive consortia. Arguably, the biggest collaborative achievements have been in generating, managing, and sharing mechanistic knowledge and data,

producing research tools, and in setting standards[5,13,246–252]. There has been some, but less, emphasis on developing better decision tools. However, provided anti-trust concerns[252,253] are managed, groups of companies that, collectively, have a large share of the pipeline in a given therapy area or technology class, could collaborate to transfer more value to organizations that invent, develop, and maintain screening and disease decision tools in the area of interest (for example, toxicology decision tools for new therapeutic modalities, or preclinical oncology decision tools). That way, the industry could get better decision tools and benefit from knowledge spill-overs without individual firms bearing the full cost of innovation.

[H1] Concluding remarks

This Perspective is an overt attempt to help institutions, comprised of people with diverse expertise and divergent interests, to improve the decision tools they apply in drug discovery and preclinical development. It has marshalled evidence from decision theory and the history of drug R&D to make the intellectual case for a focus on predictive validity. It has introduced financial frameworks to assign dollar values to better decision performance, to help make the commercial case. It has set out practical methods for evaluating decision tools and reflecting those evaluations in R&D investment decisions. It has also described some of the economic barriers to private sector investment in better decision tools. Few, if any, of the individual ideas in the Perspective are new. However, we do think the conjunction of ideas is new. We hope the conjunction will help to shift predictive validity in the right direction.

References

1. Morgan, P. *et al.* Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature Reviews Drug Discovery* vol. 17 167–181 (2018).
2. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery* vol. 12 581–594 (2013).
3. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: A five-dimensional framework. *Nature Reviews Drug Discovery* vol. 13 419–431 (2014).
4. Emmerich, C. H. *et al.* Improving target assessment in biomedical research: the GOT-IT recommendations. *Nature Reviews Drug Discovery* vol. 20 64–81 (2021).
5. Edwards, A. M. *et al.* Preclinical target validation using patient-derived cells. *Nature Reviews Drug Discovery* vol. 14 149–150 (2015).
6. Scannell, J. W. & Bosley, J. When quality beats quantity: Decision theory, drug discovery, and the reproducibility crisis. *PLoS One* **11**, (2016).
7. Paul, S. M. *et al.* How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* vol. 9 203–214 (2010).
8. Bender, A. & Cortes-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discovery Today* vol. 26 1040–1052 (2021).
9. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today* vol. 26 511–524 (2021).
10. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, (2018).
11. Bloom, N., Jones, C. I., van Reenen, J. & Webb, M. Are ideas getting harder to find?†. *Am. Econ. Rev.* **110**, 1104–1144 (2020).
12. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery* vol. 11 191–200 (2012).
13. Woodcock, J. & Woosley, R. The FDA critical path initiative and its influence on new drug development. *Annual Review of Medicine* vol. 59 1–12 (2008).
14. Horrobin, D. F. Modern biomedical research: An internally self-consistent universe with little contact with medical reality? *Nat. Rev. Drug Discov.* **2**, 151–154 (2003).
15. Steward, F. & Wibberley, G. Drug innovation - What's slowing it down? *Nature* **284**, 118–120

(1980).

16. Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug Discovery* vol. 8 959–968 (2009).
17. Le Fanu, J. *The Rise and Fall of Modern Medicine*. (Little Brown, 1999).
18. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* **10**, 428–438 (2011).
19. Gordian, M., Singh, N., Zemmel, R. & Elias, T. Why products fail in phase III. *In Vivo (Brooklyn).* **24**, 49–56 (2006).
20. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 711–715 (2004).
21. Duncan, W. A. M. & Parsons, M. E. Reminiscences of the development of cimetidine. *Gastroenterology* **78**, 620–625 (1980).
22. Shih, H. P., Zhang, X. & Aronov, A. M. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nat. Rev. Drug Discov.* **17**, 19–33 (2018).
23. Paull, K., Hodes, L. & Simon, R. M. Efficiency of antitumor screening relative to activity criteria. *J. Natl. Cancer Inst.* **76**, 1137–1142 (1986).
24. Chabner, B. A. NCI-60 Cell Line Screening: A Radical Departure in Its Time. *Journal of the National Cancer Institute* vol. 108 (2016).
25. Dykes, D. J. & Waud, W. R. Murine L1210 and P388 Leukemias. in *Tumor Models in Cancer Research* (ed. Teicher, B. A.) 23–40 (Humana Press, 2002). doi:10.1007/978-1-59259-100-8_2.
26. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).
27. DiMasi, J. A. Risks in new drug development: Approval success rates for investigational drugs. *Clinical Pharmacology and Therapeutics* vol. 69 297–307 (2001).
28. Baillar, J. C. & Gornik, H. L. Cancer Undefeated. *N. Engl. J. Med.* **336**, 1569–1574 (1997).
29. Morgan, G. W., Ward, R. & Barton, M. The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies. *Clin. Oncol.* **16**, 549–560 (2004).
30. Cairns, J. The treatment of diseases and the war against cancer. *Sci. Am.* **253**, 51–59 (1985).
31. Leaf, C. *The Truth in Small Doses: Why We're Losing the War on Cancer-and How to Win It*. (Simon & Shuster, 2013).
32. Baillar, J. C. & Smith, E. M. Progress against cancer? *N. Engl. J. Med.* **314**, 1226–1232 (1986).
33. Schilsky, R. L. & Schnipper, L. E. Hans Christian Andersen and the Value of New Cancer Treatments. *JNCI J. Natl. Cancer Inst.* **110**, 441–442 (2017).
34. Marquart, J., Chen, E. Y. & Prasad, V. Estimation of the percentage of US patients with cancer

who benefit from genome-driven oncology. *JAMA Oncol.* **4**, 1093–1098 (2018).

35. Lewis, K. Platforms for antibiotic discovery. *Nature Reviews Drug Discovery* vol. 12 371–387 (2013).
36. Bentley, R. Different roads to discovery; Prontosil (hence sulfa drugs) and penicillin (hence β-lactams). *Journal of Industrial Microbiology and Biotechnology* vol. 36 775–786 (2009).
37. da Cunha, B. R., Fonseca, L. P. & Calado, C. R. C. Antibiotic discovery: Where have we come from, where do we go? *Antibiotics* vol. 8 (2019).
38. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nature Reviews Drug Discovery* vol. 6 29–40 (2007).
39. Silver, L. L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* **24**, 71–109 (2011).
40. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* vol. 529 336–343 (2016).
41. Tommasi, R., Brown, D. G., Walkup, G. K., Manchester, J. I. & Miller, A. A. ESKAPEing the labyrinth of antibacterial discovery. *Nature Reviews Drug Discovery* vol. 14 529–542 (2015).
42. Lohmann, V. & Bartenschlager, R. On the history of hepatitis C virus cell culture systems. *Journal of Medicinal Chemistry* vol. 57 1627–1642 (2014).
43. Bartenschlager, R. Hepatitis C virus replicons: Potential role for drug development. *Nat. Rev. Drug Discov.* **1**, 911–916 (2002).
44. Lohmann, V. Hepatitis C virus cell culture models: an encomium on basic research paving the road to therapy development. *Medical Microbiology and Immunology* vol. 208 3–24 (2019).
45. Meanwell, N. A. 2015 Philip S. Portoghese medicinal chemistry lectureship. Curing hepatitis C Virus infection with direct-acting antiviral agents: The arc of a medicinal chemistry triumph. *Journal of Medicinal Chemistry* vol. 59 7311–7351 (2016).
46. Ringel, M. S., Scannell, J. W., Baedeker, M. & Schulze, U. Breaking Eroom's Law. *Nature reviews. Drug discovery* vol. 19 833–834 (2020).
47. Wu, S. S. *et al.* Reviving an R&D pipeline: a step change in the Phase II success rate. *Drug Discovery Today* vol. 26 308–314 (2021).
48. Morgan, P. *et al.* Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. *Drug Discovery Today* vol. 17 (2012).
49. Spiegelhalter, D. J., Abrams, K. R. & Myles, J. P. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* (2004). doi:10.1002/0470092602.
50. O'Hagan, A., Stevens, J. W. & Campbell, M. J. Assurance in clinical trial design. *Pharm. Stat.* **4**,

187–201 (2005).

51. Mauro, G. W., di Scala, L., Bretz, F. & Racine-Poon, A. Predictive probability of success in clinical drug development. *Epidemiol. Biostat. Public Heal.* **10**, (2013).
52. Senn, S. *Statistical Issues in Drug Development: Second Edition*. *Statistical Issues in Drug Development: Second Edition* (2008). doi:10.1002/9780470723586.
53. Willan, A. R. & Pinto, E. M. The value of information and optimal clinical trial design. *Stat. Med.* **24**, 1791–1806 (2005).
54. Bacchetti, P., McCulloch, C. E. & Segal, M. R. Simple, defensible sample sizes based on cost efficiency. *Biometrics* **64**, 577–585 (2008).
55. Bacchetti, P., Deeks, S. G. & McCune, J. M. Breaking free of sample size dogma to perform innovative translational research. *Science Translational Medicine* vol. 3 (2011).
56. Detsky, A. S. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Stat. Med.* **9**, 173–184 (1990).
57. Berry, D. A. A guide to drug discovery: Bayesian clinical trials. *Nature Reviews Drug Discovery* vol. 5 27–36 (2006).
58. Leach, A. R. & Gillet, V. J. *An introduction to chemoinformatics*. *An Introduction To Chemoinformatics* (2007). doi:10.1007/978-1-4020-6291-9.
59. Ajay, Walters, W. P. & Murcko, M. A. Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.* **41**, 3314–3324 (1998).
60. Sadowski, J. & Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325–3329 (1998).
61. Zhang, J. H., Chung, T. D. Y. & Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **4**, 67–73 (1999).
62. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* vol. 23 3–25 (1997).
63. Lipinski, C. A. Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. *Advanced Drug Delivery Reviews* vol. 101 34–41 (2016).
64. Walters, W. P. & Namchuk, M. Designing screens: How to make your hits a hit. *Nature Reviews Drug Discovery* vol. 2 259–266 (2003).
65. Bender, A. *et al.* Which aspects of HTS are empirically correlated with downstream success? *Current Opinion in Drug Discovery and Development* vol. 11 327–337 (2008).

66. Bickerton, G. R., Paolini, G. V, Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
67. Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M. & Chen, H. Chemical predictive modelling to improve compound quality. *Nature Reviews Drug Discovery* vol. 12 948–962 (2013).
68. Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C. & Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discov.* **13**, 105–121 (2014).
69. Drug Baron. Monte Carlo models of drug R&D focus attention on cutting costs – Part 1. *DRUGBARON BLOG* https://drugbaron.com/monte-carlo-models-of-drug-rd-focus-attention-on-cutting-costs-part-1/ (2013).
70. Peck, R. W., Lendrem, D. W., Grant, I., Lendrem, B. C. & Isaacs, J. D. Why is it hard to terminate failing projects in pharmaceutical R&D? *Nature Reviews Drug Discovery* vol. 14 663–664 (2015).
71. Satopää, V. A., Salikhov, M., Tetlock, P. E. & Mellers, B. Bias, Information, Noise: The BIN Model of Forecasting. *Manage. Sci.* (2021) doi:10.1287/mnsc.2020.3882.
72. Hinkle, D. E., Wiersma, W. & Jurs, S. G. *Applied Statistics for the Behavioral Sciences*. (Houghton Mifflin, 2003).
73. Akoglu, H. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* vol. 18 91–93 (2018).
74. StatisticsSolutions. Pearson's Correlation Coefficient. https://www.statisticssolutions.com/pearsons-correlation-coefficient/ (2021).
75. NCSS. Confidence intervals for Pearson's correlation. *PASS Sample Size Software* https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_Pearsons_Correlation.pdf.
76. Plenge, R. M. Disciplined approach to drug discovery and early development. *Sci. Transl. Med.* **8**, 349ps15 (2016).
77. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).
78. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, (2019).
79. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

80. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci. Rep.* **9**, 18911 (2019).
81. Sittampalam, G. *et al. Assay Guidance Manual*. *Assay Guidance Manual* (2016).
82. Williams, M., Mullane, K. & Curtis, M. J. Addressing Reproducibility: Peer Review, Impact Factors, Checklists, Guidelines, and Reproducibility Initiatives. in *Research in the Biomedical Sciences: Transparent and Reproducible* (eds. Williams, M., Curtis, M. J. & Mullane, K.) 197–306 (Academic Press, 2018). doi:10.1016/B978-0-12-804725-5.00005-7.
83. Mullane, K. & Williams, M. Enhancing reproducibility: Failures from Reproducibility Initiatives underline core challenges. *Biochemical Pharmacology* vol. 138 7–18 (2017).
84. Vollert, J. *et al.* Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Sci.* **4**, e100046 (2020).
85. Perrin, S. Make mouse studies work. *Nature* vol. 507 423–425 (2014).
86. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, (2015).
87. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. Improving bioscience research reporting: The arrive guidelines for reporting animal research. *PLoS Biol.* **8**, (2010).
88. Snyder, H. M. *et al.* Guidelines to improve animal study design and reproducibility for Alzheimer's disease and related dementias: For funders and researchers. *Alzheimer's and Dementia* vol. 12 1177–1185 (2016).
89. Lapchak, P. A., Zhang, J. H. & Noble-Haeusslein, L. J. RIGOR Guidelines: Escalating STAIR and STEPS for Effective Translational Research. *Transl. Stroke Res.* **4**, 279–285 (2013).
90. Hooijmans, C. R. *et al.* SYRCLE's risk of bias tool for animal studies. *BMC Med. Res. Methodol.* **14**, 1–9 (2014).
91. Macleod, M. R. *et al.* Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. *PLoS Biol.* **13**, 1–12 (2015).
92. Horvath, P. *et al.* Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov.* **15**, 751–769 (2016).
93. Belzung, C. & Lemoine, M. Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. *Biol. Mood Anxiety Disord.* **1**, (2011).
94. Colquhoun, D. An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. Open Sci.* **1**, (2014).
95. Ioannidis, J. P. A. Why Most Published Research Findings Are False Modeling the Framework for False Positive Findings. *PLoS Med* **2**, (2005).

96. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
97. Chuang-Stein, C. *et al.* A Quantitative Approach for Making Go/No-Go Decisions in Drug Development. *Ther. Innov. Regul. Sci.* **45**, 187–202 (2011).
98. Ferreira, G. S. *et al.* A standardised framework to identify optimal animal models for efficacy assessment in drug development. *PLoS One* **14**, (2019).
99. Veening-Griffioen, D. H. *et al.* Tradition, not science, is the basis of animal model selection in translational and applied research. *ALTEX* **38**, 49–62 (2021).
100. Prescod-Weinstein, C. What does dark matter even do? *New Sci.* **247**, 24 (2020).
101. Hubbard, D. *How to measure anything: finding the value of 'intangibles' in business*. (Wiley, 2014).
102. Funtowicz, S. O. & Ravetz, J. R. *Uncertainty and Quality in Science for Policy*. *Uncertainty and Quality in Science for Policy* (Kluwer Academic Publishers, 1990). doi:10.1007/978-94-009-0621-1.
103. Dias, L., Morton, A. & Quigley, J. *Elicitation. The Science and Art of Structuring Judgement*. *International Series in Operations Research and Management Science* vol. 261 (Springer, 2018).
104. Tetlock, Philip E., G. D. *Superforecasting : the art and science of prediction*. (Random House, 2015).
105. Baudy, A. R. *et al.* Liver microphysiological systems development guidelines for safety risk assessment in the pharmaceutical industry. *Lab on a Chip* vol. 20 215–225 (2020).
106. Gurusamy, K. S. *et al.* Clinical relevance assessment of animal preclinical research (RAA) tool: development and explanation. *PeerJ* **9**, e10673 (2021).
107. Collins, A., Ross, J. & Lang, S. H. A systematic review of the asymmetric inheritance of cellular organelles in eukaryotes: A critique of basic science validity and imprecision. *PLoS ONE* vol. 12 (2017).
108. Ekert, J. E. *et al.* Recommended Guidelines for Developing, Qualifying, and Implementing Complex In Vitro Models (CIVMs) for Drug Discovery. *SLAS Discov.* **25**, 1174–1190 (2020).
109. Friedrich, C. M. A model qualification method for mechanistic physiological QSP models to support model-informed drug development. *CPT Pharmacometrics Syst. Pharmacol.* **5**, 43–53 (2016).
110. Wehling, M. Assessing the translatability of drug projects: What needs to be scored to predict success? *Nat. Rev. Drug Discov.* **8**, 541–546 (2009).
111. Collins, A. T. & Lang, S. H. A systematic review of the validity of patient derived xenograft (PDX) models: The implications for translational research and personalised medicine. *PeerJ*

2018, e5981 (2018).

112. Willner, P. The validity of animal models of depression. *Psychopharmacology* vol. 83 1–16 (1984).
113. Morgan, M. G. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America* vol. 111 7176–7184 (2014).
114. Aspinall, W. A route to more tractable expert advice. *Nature* vol. 463 294–295 (2010).
115. Cooke, R. M. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. (Oxford University Press, 1991).
116. Chalmers, J. & Armour, M. The delphi technique. in *Handbook of Research Methods in Health Social Sciences* 715–735 (2019). doi:10.1007/978-981-10-5251-4_99.
117. Kahneman, D., Slovic, P. & Tversky, A. *Judgment under Uncertainty: Heuristics and Biases*. (Cambridge University Press, 1982).
118. Katsagounos, I., Thomakos, D. D., Litsiou, K. & Nikolopoulos, K. Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *Eur. J. Oper. Res.* **289**, 107–117 (2021).
119. Mellers, B. *et al.* Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspect. Psychol. Sci.* **10**, 267–281 (2015).
120. Bar-Hillel, M. The base-rate fallacy in probability judgments. *Acta Psychol. (Amst).* **44**, 211–233 (1980).
121. Murphy, A. H. & Daan, H. Impacts of Feedback and Experience on the Quality of Subjective Probability Forecasts. Comparison of Results from the First and Second Years of the Zierikzee Experiment. *Mon. Weather Rev.* **112**, 413–423 (1984).
122. Montibeller, G. & von Winterfeldt, D. Individual and group biases in value and uncertainty judgments. in *International Series in Operations Research and Management Science* (eds. Dias, L., Morton, A. & Quigley, J.) vol. 261 377–392 (Springer, 2018).
123. Gawande, A. *The Checklist Manifesto*. (Henry Holt and Company, 2009).
124. Kleinmuntz, D. N. *Decomposition and the control of erros in decision analytic models*. (1988).
125. Henrion, M., Fischer, G. W. & Mullin, T. Divide and conquer? effects of decomposition on the accuracy and calibration of subjective probability distributions. *Organ. Behav. Hum. Decis. Process.* **55**, 207–227 (1993).
126. Andradóttir, S. & Bier, V. M. An analysis of decomposition for subjective estimation in decision analysis. *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans.* **28**, 443–453 (1998).
127. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research.

Nature **483**, 531–533 (2012).

128. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* vol. 10 712–713 (2011).
129. Perel, P. *et al.* Comparison of treatment effects between animal experiments and clinical trials: Systematic review. *Br. Med. J.* **334**, 197–200 (2007).
130. Sena, E., Wheble, P., Sandercock, P. & Macleod, M. Systematic review and meta-analysis of the efficacy of tirilazad in experimental stroke. *Stroke* vol. 38 388–394 (2007).
131. Howells, D. W., Sena, E. S. & Macleod, M. R. Bringing rigour to translational medicine. *Nature Reviews Neurology* (2014) doi:10.1038/nrneurol.2013.232.
132. Atkinson, M. A. Evaluating preclinical efficacy. *Science Translational Medicine* vol. 3 96cm22 (2011).
133. Mullane, K. & Williams, M. Enhancing reproducibility: Failures from Reproducibility Initiatives underline core challenges. *Biochem. Pharmacol.* **138**, 7–18 (2017).
134. Reichlin, T. S., Vogt, L. & Würbel, H. The researchers' view of scientific rigor-survey on the conduct and reporting of in vivo research. *PLoS One* **11**, (2016).
135. Landis, S. C. *et al.* A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* vol. 490 187–191 (2012).
136. Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A. & Brattelid, T. PREPARE: guidelines for planning animal research and testing. *Lab. Anim.* **52**, 135–141 (2018).
137. Wendler, A. & Wehling, M. Translatability score revisited: differentiation for distinct disease areas. *J. Transl. Med.* **15**, 226 (2017).
138. Wendler, A. & Wehling, M. Translatability scoring in drug development: Eight case studies. *J. Transl. Med.* (2012) doi:10.1186/1479-5876-10-39.
139. Voelkl, B., Vogt, L., Sena, E. S. & Würbel, H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biol.* **16**, (2018).
140. Bodden, C. *et al.* Heterogenising study samples across testing time improves reproducibility of behavioural data. *Sci. Rep.* **9**, (2019).
141. Helene Richter, S. Systematic heterogenization for better reproducibility in animal experimentation. *Lab Animal* vol. 46 343–349 (2017).
142. Ferreira, G. S. Tools to enable animal to human translation: assessing the value of disease models. (Utrecht University, 2021). doi:https://doi.org/10.33540/157.
143. Okamoto, K. & Aoki, K. Development of a strain of spontaneously hypertensive rats. *Japanese J. Circ.* **27**, 282–293 (1963).
144. Veening-Griffioen, D. H. *et al.* Are some animal models more equal than others? A case study

on the translational value of animal models of efficacy for Alzheimer's disease. *Eur. J. Pharmacol.* **859**, (2019).

145. Kahneman, D. & Tversky, A. On the psychology of prediction. *Psychol. Rev.* **80**, 237–251 (1973).
146. Fabre, K. *et al.* Introduction to a manuscript series on the characterization and use of microphysiological systems (MPS) in pharmaceutical safety and ADME applications. *Lab on a Chip* vol. 20 1049–1057 (2020).
147. Proctor, W. R. *et al.* Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. *Arch. Toxicol.* **91**, 2849–2863 (2017).
148. Tetlock, P. E. *Expert political judgment: How good is it? How can we know?* (Princeton University Press, 2005). doi:10.2307/j.ctt7spbt.
149. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **10**, 507–519 (2011).
150. Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: Origins and evolution. *Nat. Rev. Drug Discov.* **13**, 577–587 (2014).
151. Beck, H. & Yee, D. Minireview: Were the IGF signaling inhibitors all bad? *Molecular Endocrinology* vol. 29 1549–1557 (2015).
152. Baserga, R. The decline and fall of the IGF-I receptor. *Journal of Cellular Physiology* vol. 228 675–679 (2013).
153. Brookes, P. The early history of the biological alkylating agents, 1918-1968. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **233**, 3–14 (1990).
154. DeVita, V. T. & Chu, E. A history of cancer chemotherapy. *Cancer Research* vol. 68 8643–8653 (2008).
155. Chabner, B. A. & Roberts, T. G. Chemotherapy and the war on cancer. *Nature Reviews Cancer* vol. 5 65–72 (2005).
156. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).
157. Varmus, H. E. The molecular genetics of cellular oncogenes. *Annual review of genetics* vol. 18 553–612 (1984).
158. Morange, M. From the Regulatory Vision of Cancer to the Oncogene Paradigm, 1975-1985. *J. Hist. Biol.* **30**, 1–29 (1997).
159. Bazell, R. *Her-2: The Making of Herceptin, a Revolutionary Treatment for Breast Cancer*. (Random House, 1998).
160. Mukherjee, S. *The Emperor of All Maladies: A Biography of Cancer*. (Scribner, 2010).

161. Weinstein, I. B. Cancer: Addiction to oncogenes - The Achilles heal of cancer. *Science* vol. 297 63–64 (2002).
162. Sawyers, C. L. Shifting paradigms: The seeds of oncogene addiction. *Nature Medicine* vol. 15 1158–1161 (2009).
163. Kersten, K., Visser, K. E., Miltenburg, M. H. & Jonkers, J. Genetically engineered mouse models in oncology research and cancer medicine. *EMBO Mol. Med.* **9**, 137–153 (2017).
164. Hidalgo, M. *et al.* Patient-derived Xenograft models: An emerging platform for translational cancer research. *Cancer Discov.* **4**, 998–1013 (2014).
165. Bhimani, J., Ball, K. & Stebbing, J. Patient-derived xenograft models—the future of personalised cancer treatment. *British Journal of Cancer* vol. 122 601–602 (2020).
166. Bonekamp, N. A. *et al.* Small-molecule inhibitors of human mitochondrial DNA transcription. *Nature* **588**, 712–716 (2020).
167. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
168. Shawver, L. K., Slamon, D. & Ullrich, A. Smart drugs: Tyrosine kinase inhibitors in cancer therapy. *Cancer Cell* vol. 1 117–123 (2002).
169. Davis, C. *et al.* 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* **359**, (2017).
170. Tannock, I. F. & Hickman, J. A. Limits to personalized cancer medicine. *N. Engl. J. Med.* (2016) doi:10.1056/NEJMsb1607705.
171. Booth, C. M. & Del Paggio, J. C. Approvals in 2016: Questioning the clinical benefit of anticancer therapies. *Nature Reviews Clinical Oncology* vol. 14 135–136 (2017).
172. Hwang, T. J. *et al.* Efficacy, safety, and regulatory approval of food and drug administration–Designated breakthrough and nonbreakthrough cancer medicines. *J. Clin. Oncol.* **36**, (2018).
173. Prasad, V. Our best weapons against cancer are not magic bullets. *Nature* vol. 577 451 (2020).
174. Haslam, A. & Prasad, V. Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs. *JAMA Netw. open* **2**, (2019).
175. Middleton, G., Robbins, H., Andre, F. & Swanton, C. A state-of-the-art review of stratified medicine in cancer: towards a future precision medicine strategy in cancer. *Annals of Oncology* vol. 33 143–157 (2022).
176. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* vol. 15 81–94 (2018).
177. Hochhaus, A. *et al.* Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid

Leukemia. *N. Engl. J. Med.* **376**, 917–927 (2017).

178. Kesselheim, A. S. & Avorn, J. The most transformative drugs of the past 25 years: A survey of physicians. *Nature Reviews Drug Discovery* vol. 12 425–431 (2013).
179. Elliott, J. *et al.* ALK inhibitors for non-small cell lung cancer: A systematic review and network meta-analysis. *PLoS One* **15**, (2020).
180. Shah, R. & Lester, J. F. Tyrosine Kinase Inhibitors for the Treatment of EGFR Mutation-Positive Non–Small-Cell Lung Cancer: A Clash of the Generations. *Clinical Lung Cancer* vol. 21 e216–e228 (2020).
181. Comen, E., Gilewski, T. A. & Norton, L. Tumor Growth Kinetics. in *Holland-Frei Cancer Medicine* (eds. Bast, R. C. et al.) 589–600 (Wiley, 2016). doi:10.1002/9781119000822.
182. Skipper, H. E. The Effects of Chemotherapy on the Kinetics of Leukemic Cell Behavior. *Cancer Res.* **25**, 1544–1550 (1965).
183. Skipper, H. E. Kinetics of mammary tumor cell growth and implications for therapy. *Cancer* **28**, 1479–1499 (1971).
184. Kerr, J. F. R., Wyllie, A. H. & Currie, A. R. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* **26**, 239–257 (1972).
185. Hickman, J. A. Apoptosis induced by anticancer drugs. *Cancer Metastasis Rev.* **11**, 121–139 (1992).
186. Strasser, A. & Vaux, D. L. Cell Death in the Origin and Treatment of Cancer. *Molecular Cell* vol. 78 1045–1054 (2020).
187. Nowell, P. C. The clonal evolution of tumor cell populations. *Science (80-.).* **194**, 23–28 (1976).
188. Loeb, L. A. Human cancers express mutator phenotypes: Origin, consequences and targeting. *Nature Reviews Cancer* vol. 11 450–457 (2011).
189. Vendramin, R., Litchfield, K. & Swanton, C. Cancer evolution: Darwin and beyond. *EMBO J.* **40**, (2021).
190. Turajlic, S. & Swanton, C. Implications of cancer evolution for drug development. *Nature Reviews Drug Discovery* vol. 16 441–442 (2017).
191. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
192. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* vol. 168 613–628 (2017).
193. Muthuswamy, S. K. Self-organization in cancer: Implications for histopathology, cancer cell biology, and metastasis. *Cancer Cell* **39**, 443–446 (2021).

194. Hill, W., Caswell, D. R. & Swanton, C. Capturing cancer evolution using genetically engineered mouse models (GEMMs). *Trends in Cell Biology* (2021) doi:10.1016/j.tcb.2021.07.003.
195. Sotillo, R., Schvartzman, J. M., Socci, N. D. & Benezra, R. Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature* **464**, 436–440 (2010).
196. Takenaka, T., Yamazaki, K., Miura, N., Mori, R. & Takeo, S. The prognostic impact of tumor volume in patients with clinical stage IA non-small cell lung cancer. *J. Thorac. Oncol.* **11**, 1074–1080 (2016).
197. Jung, J., Seol, H. S. & Chang, S. The generation and application of patient-derived xenograft model for cancer research. *Cancer Research and Treatment* vol. 50 1–10 (2018).
198. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).
199. Ben-David, U. *et al.* Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.* **49**, 1567–1575 (2017).
200. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* vol. 481 306–313 (2012).
201. EMA. *ICH Guideline S9 on Nonclinical Evalution for Anticancer Pharmaceuticals*. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-guideline-s9-non-clinical-evaluation-anticancer-pharmaceuticals-step-5_en.pdf (2010).
202. FDA. *Guidance for Industry: S9 Nonclinical Evaluation for Anticancer Pharmaceuticals*. https://www.fda.gov/media/73161/download (2010).
203. Wieschowski, S. *et al.* Preclinical efficacy studies in investigator brochures: Do they enable risk–benefit assessment? *PLoS Biol.* **16**, (2018).
204. Hickman, J. A. *et al.* Three-dimensional models of cancer for pharmacology and cancer cell biology: Capturing tumor complexity in vitro/ex vivo. *Biotechnology Journal* vol. 9 1115–1128 (2014).
205. Misra, S. *et al.* Ex vivo organotypic culture system of precision-cut slices of human pancreatic ductal adenocarcinoma. *Sci. Rep.* **9**, 2133 (2019).
206. Sontheimer-Phelps, A., Hassell, B. A. & Ingber, D. E. Modelling cancer in microfluidic human organs-on-chips. *Nature Reviews Cancer* vol. 19 65–81 (2019).
207. Burian, R. M. How the choice of experimental organism matters: Epistemological reflections on an aspect of biological practice. *J. Hist. Biol.* **26**, 351–367 (1993).
208. Vorrink, S. U., Zhou, Y., Ingelman-Sundberg, M. & Lauschke, V. M. Prediction of drug-induced hepatotoxicity using long-term stable primary hepatic 3D spheroid cultures in chemically defined conditions. *Toxicol. Sci.* **163**, 655–665 (2018).

209. Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J. & Huang, R. S. Consistency in large pharmacogenomic studies. *Nature* vol. 540 E1–E2 (2016).
210. Haibe-Kains, B. *et al.* Revisiting inconsistency in large pharmacogenomic studies. *F1000Research* **5**, (2017).
211. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–393 (2013).
212. Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* (2016) doi:10.1038/nature17987.
213. Mayr, L. M. & Bojanic, D. Novel trends in high-throughput screening. *Current Opinion in Pharmacology* vol. 9 580–588 (2009).
214. Pedró-Rosa, L. *et al.* Identification of potent inhibitors of the trypanosoma brucei methionyl-tRNA synthetase via high-throughput orthogonal screening. *J. Biomol. Screen.* **20**, 122–130 (2015).
215. Mondritzki, T. Prädiktive Wertigkeit verschiedener präklinischer Outcome-Parameter für eine erfolgreiche versus nicht-erfolgreiche klinische Entwicklung von Arzneimitteln zur Behandlung der Herzinsuffizienz. (Donau-Universität-Krems, 2014).
216. Moser, J. & Verdin, P. Burgeoning oncology pipeline raises questions about sustainability. *Nature Reviews Drug Discovery* vol. 17 698–699 (2018).
217. Deloitte. *Ten years on. Measuring the return from pharmaceutical innovation*. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-ten-years-on-measuring-return-on-pharma-innovation-report-2019.pdf (2019).
218. Skardal, A. *et al.* Multi-tissue interactions in an integrated three-tissue organ-on-a-chip platform. *Sci. Rep.* **7**, (2017).
219. Maurer, T. Model-Based Discovery and Development of Novel Therapies for Type-2 Diabetes Mellitus. in *Bridging Bench and Bedside with Quantitative Model-Based Translational Pharmacology* (New York Academy of Science, 2012).
220. Bertau, M., Mosekilde, E. & Westerhoff, H. V. *Biosimulation in Drug Development*. *Biosimulation in Drug Development* (Wiley-VCH, 2008). doi:10.1002/9783527622672.
221. Mager, D. E. & Kimko, H. H. C. *Systems Pharmacology and Pharmacodynamics. AAPS Advances in the Pharmaceutical Sciences Series*. (Springer International Publishing, 2016). doi:10.1007/978-3-319-44534-2.
222. Helmlinger, G. *et al.* Quantitative Systems Pharmacology: An Exemplar Model-Building Workflow With Applications in Cardiovascular, Metabolic, and Oncology Drug Development.

CPT: Pharmacometrics and Systems Pharmacology vol. 8 380–395 (2019).

223. Mardinoglu, A. *et al.* The potential use of metabolic cofactors in treatment of NAFLD. *Nutrients* vol. 11 1578 (2019).
224. Jørgensen, P. G. *et al.* Cardiac adaptation in hibernating, free-ranging Scandinavian Brown Bears (Ursus arctos). *Sci. Rep.* **10**, (2020).
225. Peretti, D. *et al.* RBM3 mediates structural plasticity and protective effects of cooling in neurodegeneration. *Nature* **518**, 236–239 (2015).
226. Krystal, A. D. *et al.* A randomized proof-of-mechanism trial applying the 'fast-fail' approach to evaluating κ-opioid antagonism as a treatment for anhedonia. *Nat. Med.* **26**, 760–768 (2020).
227. Grabb, M. C., Hillefors, M. & Potter, W. Z. The NIMH 'Fast-Fail Trials' (FAST) Initiative: Rationale, Promise, and Progress. *Pharmaceut. Med.* **34**, 233–245 (2020).
228. Krystal, A. D. *et al.* The first implementation of the NIMH FAST-FAIL approach to psychiatric drug development. *Nature Reviews Drug Discovery* vol. 18 82–84 (2018).
229. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Med.* **11**, (2013).
230. Insel, T. *et al.* Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry* vol. 167 748–751 (2010).
231. Cuthbert, B. N. The PRISM project: Social withdrawal from an RDoC perspective. *Neuroscience and Biobehavioral Reviews* vol. 97 34–37 (2019).
232. Sullivan, P. F. *et al.* A mega-Analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
233. Healy, D. *The Antidepressant Era*. (Harvard University Press, 1997).
234. Liu, K. S. *et al.* Is bigger better for depression trials? *J. Psychiatr. Res.* **42**, 622–630 (2008).
235. FDA. *Major Depressive Disorder: Developing Drugs for Treatment Guidance for Industry*. (2018).
236. Nelson, R. R. The Simple Economics of Basic Scientific Research. *J. Polit. Econ.* **67**, 297–306 (1959).
237. Arrow, K. J. Economic Welfare and the Allocation of Resources for Invention. in *Readings in Industrial Economics* (ed. Rowley, C. K.) 219–236 (Palgrave, 1972). doi:10.1007/978-1-349-15486-9_13.
238. Dobson, C. M. Chemical space and biology. *Nature* vol. 432 824–828 (2004).
239. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* vol. 432 855–861 (2004).
240. Seymore, S. B. Making patents useful. *Minn. Law Rev.* **98**, 1046–1109 (2014).

241. Billette de Villemeur, E. & Versaevel, B. One lab, two firms, many possibilities: On R&D outsourcing in the biopharmaceutical industry. *J. Health Econ.* **65**, 260–283 (2019).
242. Hoofnagle, J. H. & Sherker, A. H. Therapy for Hepatitis C — The Costs of Success. *N. Engl. J. Med.* **370**, 1552–1553 (2014).
243. Senate Committee Is Investigating Pricing of Hepatitis C Drug. *Wall Street Journal* https://www.wsj.com/articles/senate-finance-committee-is-investigating-pricing-of-hepatitis-c-drug-1405109206 (2014).
244. Morrison, C. 2019 biotech IPOs: party on. *Nature reviews. Drug discovery* vol. 19 6–9 (2020).
245. Morrison, C. Boom: 2018's biotech IPOs. *Nat. Rev. Drug Discov.* **18**, 3–6 (2018).
246. Williamson, A. R. Creating a structural genomics consortium. *Nature Structural Biology* vol. 7 953 (2000).
247. Vaudano, E. The Innovative Medicines Initiative: A public private partnership model to foster drug discovery. *Comput. Struct. Biotechnol. J.* **6**, e201303017 (2013).
248. Ochoa, D. *et al.* Open Targets Platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).
249. Dolgin, E. Massive NIH–industry project opens portals to target validation. *Nat. Rev. Drug Discov.* **18**, 240–242 (2019).
250. Holden, A. L. The SNP Consortium: Summary of a private consortium effort to develop an applied map of the human genome. *BioTechniques* vol. 32 S22–S26 (2002).
251. Holden, A. L., Contreras, J. L., John, S. & Nelson, M. R. The international serious adverse events consortium. *Nature Reviews Drug Discovery* vol. 13 795–796 (2014).
252. Contreras, J. L. & Vertinsky, L. S. Pre-Competition. *North Carol. Law Rev.* **95**, 67–131 (2016).
253. Lundqvist, B. Joint Research and Development Collaborations Under Competition Law, with a Layman's Economic Viewpoint. *SSRN Electron. J.* (2017) doi:10.2139/ssrn.2913840.
254. Brown, D. G., May-Dracka, T. L., Gagnon, M. M. & Tommasi, R. Trends and exceptions of physical properties on antibacterial activity for gram-positive and gram-negative pathogens. *J. Med. Chem.* **57**, 10144–10161 (2014).
255. Zoffmann, S. *et al.* Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **9**, (2019).
256. Farha, M. A. & Brown, E. D. Unconventional screening approaches for antibiotic discovery. *Ann. N. Y. Acad. Sci.* **1354**, 54–66 (2015).
257. Yokokawa, F. Recent Progress on the Development of Novel Antitubercular Agents from Whole-Cell Screening Hits. *J. Synth. Org. Chem. Japan* **72**, 1239–1249 (2014).
258. Horscroft, N. *et al.* Replicon cell culture system as a valuable tool in antiviral drug discovery

against hepatitis C virus. *Antiviral Chemistry and Chemotherapy* vol. 16 1–12 (2005).

259. Woerz, I., Lohmann, V. & Bartenschlager, R. Hepatitis C virus replicons: Dinosaurs still in business? *Journal of Viral Hepatitis* vol. 16 1–9 (2009).
260. Kaplan, G. & Racaniello, V. R. Construction and characterization of poliovirus subgenomic replicons. *J. Virol.* **62**, 1687–1696 (1988).
261. Khromykh, A. A. & Westaway, E. G. Subgenomic replicons of the flavivirus Kunjin: construction and applications. *J. Virol.* **71**, 1497–1505 (1997).
262. Behrens, S.-E., Grassmann, C. W., Thiel, H.-J., Meyers, G. & Tautz, N. Characterization of an Autonomous Subgenomic Pestivirus RNA Replicon. *J. Virol.* **72**, 2364–2372 (1998).
263. Lohmann, V. *et al.* Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science (80-.).* **285**, 110–113 (1999).
264. Chung, R. T. & Baumert, T. F. Curing Chronic Hepatitis C — The Arc of a Medical Triumph. *N. Engl. J. Med.* **370**, 1576–1578 (2014).
265. Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R. & Howells, D. W. Systematic reviews and meta-analysis of preclinical studies: Why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow and Metabolism* vol. 34 737–742 (2014).
266. Macleod, M. R., O'Collins, T., Howells, D. W. & Donnan, G. A. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke* vol. 35 1203–1208 (2004).
267. MacLeod, M. R. *et al.* Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* **39**, 2824–2829 (2008).
268. Macleod, M. R. *et al.* Good laboratory practice: preventing introduction of bias at the bench. *Stroke.* **40**, (2009).
269. Dirnagl, U., Iadecola, C. & Moskowitz, M. A. Pathobiology of ischaemic stroke: An integrated view. *Trends in Neurosciences* vol. 22 391–397 (1999).
270. Dronne, M. A., Grenier, E., Chapuisat, G., Hommel, M. & Boissel, J. P. A modelling approach to explore some hypotheses of the failure of neuroprotective trials in ischemic stroke patients. *Progress in Biophysics and Molecular Biology* vol. 97 60–78 (2008).
271. Choi, D. W. Excitotoxicity: Still Hammering the Ischemic Brain in 2020. *Frontiers in Neuroscience* vol. 14 (2020).
272. Orset, C. *et al.* Efficacy of Alteplase in a Mouse Model of Acute Ischemic Stroke: A Retrospective Pooled Analysis. *Stroke* **47**, 1312–1318 (2016).
273. Phipps, M. S. & Cronin, C. A. Management of acute ischemic stroke. *The BMJ* vol. 368 l6983 (2020).
274. Sams-Dodd, F. Strategies to optimize the validity of disease models in the drug discovery

process. *Drug Discovery Today* vol. 11 355–363 (2006).

275. Hooijmans, C. R., De Vries, R., Leenaars, M., Curfs, J. & Ritskes-Hoitinga, M. Improving planning, design, reporting and scientific quality of animal experiments by using the Gold Standard Publication Checklist, in addition to the ARRIVE guidelines. *British Journal of Pharmacology* vol. 162 1259–1260 (2011).
276. Fisher, M. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* **30**, 2752–2758 (1999).
277. Vettoretti, M., Facchinetti, A., Sparacino, G. & Cobelli, C. Predicting Insulin Treatment Scenarios with the Net Effect Method: Domain of Validity. *Diabetes Technol. Ther.* **18**, 694–704 (2016).
278. Sacca, L., Toffolo, G. & Cobelli, C. V-A and A-V modes in whole body and regional kinetics: Domain of validity from a physiological model. *American Journal of Physiology - Endocrinology and Metabolism* vol. 263 E597–E606 (1992).
279. Rudolf, A. F., Skovgaard, T., Knapp, S., Jensen, L. J. & Berthelsen, J. A comparison of protein kinases inhibitor screening methods using both enzymatic activity and binding affinity determination. *PLoS One* **9**, (2014).

Related links

Sample size calculator from Statistics Kingdom: https://www.statskingdom.com/sample_size_regression.html

The Collaborative Adverse Outcome Pathway Wiki: https://aopwiki.org/

DataFAIRy Bioassay Annotation: https://www.pistoiaalliance.org/projects/current-projects/datafairy-bioassay-annotation/

Springer Nature Experiments: https://experiments.springernature.com/

Acknowledgements

We thank: Professor Alec Morton and Dr Abigail Colson at the Department of Management Science at Strathclyde University for their help on the decision tool evaluation section; Professor Bruno Versaevel and Professor Etienne Billet de Villemeur, at EMLYON Business School and at Université des Sciences et Technologies de Lille, respectively, for their help on the economics of decision tool-related innovation; and Dr Lorna Ewart, EVP Science at Emulate, for her help with toxicology-related content. We also thank Professor Matt Todd at the UCL School of Pharmacy for helpful comments on an earlier draft of the paper.

Competing interests

J.W.S. is a director and shareholder of JW Scannell Analytics, which sells consulting services to the biopharmaceutical and financial services sectors, including to firms that are commercialising screening and disease models. He is an employee of Unify Pharmaceuticals. He holds equity options in Ochre Bio. J. B. is an employee of Novadiscovery and has equity options in the firm. J.A. Hickman has no competing interests. G.R.D. is a director and major shareholder of P1vital and P1vital Products Ltd. P1vital provides clinical research services to the pharmaceutical industry. P1vital products provides data management and digital technology to the pharmaceutical industry and healthcare providers. Hubert Truebel owns shares in Bayer AG. G.S.F. has received fees from Merck KGaA and Curare Consulting. He has been employed by the Access to Medicine Foundation and 3D PharmXchange. He is currently employed by Janssen Biologics. J.M.T.is a director and/or shareholder of Talisman Therapeutics, Gen2 Neuroscience, Avilex Pharma, Domain Therapeutics, Cellesce and Ubiquigent.

List of supplementary material

1. Q&A on the practical relevance of the decision theoretic modelling.
2. Mathematica code to generate the decision-theoretic analyses for Figure 2 and Figure 3.
3. Excel spreadsheet to compute the financial analyses for Figure 3 and Figure 4.
4. Education and awareness: A validity-related reading list with brief notes.
5. Commentary on phenotypic versus target-based drug discovery, and serendipitous discovery.

Tables

Table 1 | Glossary of terms and comments

Term	Definition	Comments
Decision tool	Tools to inform R&D decisions are used to evaluate therapeutic candidates by generating measures that are believed to correlate with clinical utility.	Our "decision tool" corresponds roughly to Sams-Dodd's[274] "screening system" and to Scannell and Bosley's[6] "predictive model".
Domains of validity	The parameter space within which the decision tool is predictive.	An important question when evaluating decision tools is whether the domains of validity include clinical utility in people. See Table 2 and Boxes 1 to 3.
Predictive validity	The degree to which the ordering of measures from decision tool would match, across a population of therapeutic candidates, the ordering in terms of clinical utility in people. We operationalise predictive validity as the notional Pearson correlation coefficient between the decision tool output and the relevant measure of clinical utility. However, it would be reasonable to operationalize predictive validity in other ways[6] (for example, Spearman's rank correlation, or area under the ROC curve).	There is a difference in detail in the way we define predictive validity versus Scannell and Bosley[6]. Scannell and Bosley excluded sampling error from their treatment to formally separate the analysis of what they called "Predictive Validity," which cannot be managed by increasing sample size, from that of reliability, which can. However, for practical R&D management, when considering whether or not to believe the results of a decision tool, or how to improve a decision tool, it makes little sense to ignore sampling error. Therefore, any de-correlating effect of sampling error would be included in the predictive validity definition we use in this paper.
True positives (*TP*)	Therapeutic candidates classified as *"yeses"* by the decision tool which are positives, having sufficient clinical utility in people.	-
True negatives (*TN*)	Therapeutic candidates classified as "*noes*" by the decision tool which are negatives, having insufficient clinical utility in people.	-
False positives (*FP*)	Therapeutic candidates classified as "*yeses*" by the decision tool which are negatives, having insufficient clinical utility in people.	-
False negatives (*FN*)	Therapeutic candidates classified as "*noes*" by the decision tool but are positives, which would have had sufficient clinical utility in people.	-
True positive rate (*TPR*)	$TPR = \#TP/(\#TP + \#FN)$	Often neglected in the discussion of drug R&D productivity but, as with PPV, very sensitive to predictive validity.
False positive rate (*FPR*)	$FPR = \#FP/(\#FP + \#TN)$	-

Positive predictive value (*PPV*)	$PPV = \#TP/(\#TP + \#FP) = 1 - FDR$.	PPV is an important measure in drug R&D because unit costs tend to rise as candidates progress through the process.
False discovery rate (*FDR*)	$FDR = \#FP/(\#TP + \#FP) = 1 - PPV$	-
Pearson correlation coefficient	Correlation = $\rho_{X,Y} = cov\,(X,Y)/\sigma_X\sigma_Y$ Where X and Y are a pair or random variables, cov is the covariance, and σ_X σ_Y are standard deviations.	This is the "standard" correlation coefficient that is most commonly used.
Spearman rank correlation coefficient	The Pearson correlation coefficient where X and Y are expressed as rank variables.	A common measure of rank correlation.
ROC curve	A graph that shows the performance of a binary classifier as the decision threshold varies. It plots the FPR of the decision (from 0 to 1) against the TPR of the decision (from 0 to 1).	First developed to measure the performance of radar receivers in World War II, hence the term Receiver Operating Characteristic (ROC) curve. Now used to measure decision performance in a wide range of disciplines.
Area under the ROC curve	The area between the ROC curve and the FPR (horizontal) axis between FPR = 0 and FPR = 1.	This measure considers decision performance across all possible decision thresholds. When decision performance is very good, TPR >> FPR, so the area under the ROC curve approaches 1. When decision performance is the same as a random guess, TPR = FPR so the area under the ROC curve is 0.5.
Net present value (NPV)	The sum of a time series of cashflows discounted to the present time at the prevailing time cost of money (for example, interest rate or required rate of return)	When NPV is positive, the programme has a value that is higher than the cost of the funds, and when NPV is negative, the cost of the funds is higher than the value of the programme
Confusion matrix	A simple 2×2 table with the structure shown below, where P = positive and N = negative	When actual positives are rare (as is generally the case in drug discovery) and predictive validity is low, false positives are common.

Commented [PK1]: MPS – please add the table image in this cell (please ignore this comment Jack – it is for our production team)

Table 2 | Four dimensions for decision tool evaluation

Examples of evaluation criteria	Overlapping terminology	Examples in this article	Selected references, resources and comments
1. Biological recapitulation			
To what extent does the decision tool resemble the human clinical state in terms of: epidemiology; symptoms and natural history; genetics: biochemistry; aetiology; histology; biomarkers; and response to known human pharmacology (including positive and negative controls)?	External validity, face validity, construct validity	Boxes 1 to 3, depression (main text), oncology (main text)	[98,105–107,111,274]
2. Tests and endpoints			
To what extent: is the experimental protocol similar to the likely clinical treatment regimen; does drug dosing and tissue exposure match the likely clinical situation; are the endpoints used in the preclinical studies translatable to the likely clinical endpoints; are the methods used to measure preclinical endpoints comparable to the likely clinical measures; is there confidence in the go/no-go thresholds that we will apply to the measures that the decision tool yields?	External validity, face validity, construct validity	Box 1, Box 2, oncology (main text)	[98,106,132,142,274]
3. Experimental and statistical hygiene			
To what extent is testing implemented with: animals derived from trusted sources; with confirmed genotype; randomized and blinded animal allocation and assessment; appropriate sample size?	Internal validity, reproducibility, replicability, bias, reporting quality	Box 3, depression (main text)	[88,89,140,141,275,276,90,91,94,106,107,111,132,139]; see also the CAMERADES collaboration, ARRIVE guidelines[87], PREPARE guidelines[136]
To what extent are: results repeatable; consistent with historical results derived from the same animal strain; robust to modest changes in experimental conditions (for example, animal strain)?			
To what extent is the statistical treatment: pre-planned; methodologically appropriate; sufficiently powerful; considering false discovery rates?			
4. Domains of validity			
For which disease states, treatment regimens and clinical endpoints is decision tool output likely to correlate well with drugs' performance in people?	Context of use	Boxes 1 to 3, oncology (main text)	The idea of models' domains of validity is common in some disciplines but is rarely used explicitly in biology and medicine. However, see refs [6,108,277,278]

Figures

Figure 1 | Pharmaceutical R&D depends on selections made using a set of decision tools. The expectation is that the decision tools score therapeutic candidates (points in the scatter graphs, arbitrary units) in a way that correlates with clinical utility in human patients, so that candidates that score well on the decision tools are enriched for those that could succeed in the clinic (red dots). Of course, these correlations are rarely measured in practice, but for empirical examples of correlations between decision tools, or between decision tools and clinical utility in patients, see references[147,279]. Typically, less expensive, higher throughput, decision tools are used early in the process to increase the number of candidates that can be tested (such as an in vitro potency assay to select candidates for animal testing shown in panel **a**) with the general assumption that these tools have lower predictive validity than the more expensive decisions tools used later in the process (such as an animal efficacy assay shown in panel **b**). Panel **c** shows the conceptual basis of our decision theoretic analysis. By representing therapeutic candidates in a "measurement space" and by applying plausible decision rules to segment the space, one can explore the parameters to which R&D performance (for example, the ability to find the red dots) is sensitive. The parameters include decision tools' throughput, the candidate selection thresholds that are applied, their predictive validity, the degree to which decision tools are statistically independent of each other, the rarity of good clinical candidates, and the underlying distributions of clinical utility and decision tool scores. For a technical discussion, see Scannell and Bosley[6], and the Supplementary information for this article.

Figure 2 | Predictive validity, throughput and R&D decision performance. Panel **a** illustrates the relationship between decision tool scores (x-axis) and clinical utility (y-axis) for a hypothetical set of therapeutic candidates. The shading (z-axis) shows the probability density of candidates. Light shades indicate a high probability of finding a candidate and dark shades indicate low probability. Here the decision tool has high predictive validity (correlation coefficient = 0.95). Suppose we use the tool to evaluate candidates to select some for the clinic. True positives (upper right quadrant) are candidates that exceed our selection threshold (that is, lie to the right of the solid vertical line) and which have sufficient clinical utility to work in people (that is, are above the dotted horizontal line). False positives (lower right quadrant) exceed the selection threshold but have insufficient clinical utility, so fail later in the process. Panel **b** repeats the illustration for a decision tool with low predictive validity (correlation coefficient = 0.5). Compared with panel **a**, for any given selection threshold, more of the probability mass lies in the false positive quadrant and less lies in the true positive quadrant. Panel **c** shows how positive predictive value (PPV = true positives/(true positives + false positives)) increases as the selection threshold using the decision tool with high predictive validity shown in panel **a** rises. ~~Of course, one needs to evaluate more candidates to find the rarer ones whose scores exceed the rising selection threshold ("Throughput required"). The decision tool has high predictive validity as in panel A and~~ W~~w~~e assume that 0.1% of potential therapeutic candidates genuinely have sufficient clinical utility to work in patients (that is, ~0.1% of the probability mass lies above the horizontal dotted line illustrated in panel **a**). Of course, one needs to evaluate more candidates to find the rarer ones whose scores exceed the rising selection threshold ("Throughput required"). With ~200 candidates ("Throughput required") we expect the best candidate to score ~15 on the decision tool, and there is a ~20% chance it would exceed the clinical utility threshold (PPV ≅ 0.2). With ~600 candidates, we expect the best candidate to score ~22 on the decision tool and there is a ~45% chance it would exceed the clinical utility threshold (PPV ≅ 0.45). Panel **d** repeats the analyses in panel **c**, using the same pool of candidates, but for the decision tool with low predictive validity as in panel **b**. PPV remains low even as the decision threshold rises because candidates that exceed the selection threshold are mainly false positives. Were we to test ~2,500 candidates and pick the one with the highest score on the decision tool, the probability it would have sufficient clinical utility is less than 5%. For details of the analysis and the associated code, see Supplementary information. For what appears to be a real-world illustration, see Box 1.

Commented [PK2]: With the shading, would it be clearer if the outermost shade (representing regions with no candidates) in both graphs was not in the same family as the rest of the graph, and potentially the same colour in both graphs (e.g. a very light grey)? It seems like this could make the point about the probability mass (presumably represented by the size of the shaded areas in blue on the left and yellow/brown on the right?) clearer

Commented [PK3]: Something that is potentially confusing about panels c and d of this figure is that the score on the decision tool goes from 0 to 10 in panels a and b, but the selection threshold in panels c and d (which could be expected to be the score using the decision tool) goes from 0 to 30. Please could you either clarify why this is or adjust the figures so that this issue is clearer.

Figure 3 | R&D productivity is very sensitive to predictive validity. a | The numbered contours show how positive predictive value (PPV = true positives/(true positives + false positives)) varies as the predictive validity and throughput of the decision tool vary. Calculation methods follow Scannell and Bosley[6]. As with Figure 2, we assume that one candidate out of 1,000 (0.1%) has sufficient clinical utility (that is, is a "true positive"). For much of the parameter space, an absolute 0.1 change in predictive validity (horizontal axis) has a bigger effect on PPV than a 10× change in throughput ($\log_{10}$ scale on the vertical axis). For example, increasing predictive validity from 0.4 to 0.5 with a throughput of 10^4 raises PPV from 5% to 10%; about the same effect as increasing throughput by a factor of 40 while holding predictive validity constant at 0.4 (highlighted with dotted arrows). Panel **b** treats all activity up to phase I as an aggregate "decision tool" that delivers a fixed number of candidates into phase I trials. We use the PPV parameter (from panel **a**, where PPV contours range from 0.3% to 33%) to vary the proportion of drug candidates entering phase I that have sufficient clinical utility in people. The contours then show the expected net present value (NPV) of lifecycle cashflows per phase I candidate in millions of US$ ($m), discounted to the time of FDA approval. NPV increases with PPV because more of the phase I candidates are eventually approved and sold. The supplementary information includes an Excel-based financial model that shows the calculation in detail. We base our analysis on the archetypal 'big pharma' R&D lifecycle set out by Paul et al.[7] that conventionally leads to a single drug approval. The clinical trial costs vary (higher with more candidates that have clinical utility), as do post-approval profits (in proportion to the number of drug approvals). The profit per approved drug is set so that the financial return on R&D investment using the Paul et al.[7] parameters — which lead to a single approval — is 11%. For comparison, the aggregate NPV of R&D costs leading up to each phase I candidate in the Paul et al. model averages around −$100 million. The NPV figures in the bottom left hand of panel **b** are lower than this figure because poor candidates enter clinical trials and incur clinical trial costs that are not offset by post-approval profits. Panel **c** takes parameters along the red dotted line in panel **b**, ("Base case") and applies a multiplier to the profit per launched drug while holding the costs of R&D constant. It then calculates lifecycle NPV per phase I candidate. An 0.1 absolute change in predictive validity has nearly as big an effect on the value of R&D as a 2× change in the profit per launch. The example of the impact of an increase in predictive validity from 0.5 to 0.6 is shown on the graph (~~dashed~~ arrows). Panel **d** shows the increase in lifecycle NPV per phase I candidate from an incremental 0.1 increase in the aggregate predictive validity of the R&D activities that deliver candidates into phase I. The baseline predictive validity, on which the incremental increase is made, is shown on the horizontal axis.

Commented [PK4]: In the drawn up figures, the shades in the top right of each graph are also blue – should they actually be in a different shade (e.g. beige/grey)?

Commented [PK5]: The dotted arrows here don't seem to be fully highlighting the desired comparison. Would it be better to have a dotted line without arrows set at around 380 on the y axis as a present, a vertical arrow from the dotted red line to the yellow line at 0.5 and an L-shaped horizontal arrow from 0.5 to 0.6 at around 0 on the y axis, then going up vertically to the horizontal dotted line at 380?

Figure 4 | Financial value of decision tool evaluation in pharmaceutical R&D. The graph is derived from points along the dotted red line base case in Figure 3**d** and illustrates the financial value of the ability to discriminate between discovery and preclinical development activities with higher or lower aggregate predictive validity with regard to clinical utility in humans. One can also regard these lines as representing the maximum amount that an economically rational actor would invest in evaluations to allow them to make the discrimination. Imagine a large drug company is evaluating two smaller firms, both potential acquisition candidates, with similar archetypical portfolios of discovery and preclinical projects; again based on the Paul et al. parameters[7]. One of the target firms used screening and disease models with higher predictive validity. The lines show the expected increase in net present value (NPV) per phase I candidate as the acquirer's ability to correctly discriminate between the target firms varies, and as the spread in predictive validity between the target firms varies (shown for 0.75 versus 0.5, 0.65 versus 0.5 and 0.55 versus 0.5). For scale, the NPV of R&D spending to deliver a phase I candidate is roughly $100 million (m), including the cost of failure. The underlying financial assumptions and analysis are the same as Figure 3 and are set out in detail in an Excel spreadsheet available in the Supplementary information.

Boxes

Commented [PK6]: In the journal format, references that only appear in display items are cited in the order: tables, figures, boxes, which means some renumbering is needed to reflect this position. Hopefully this is straightforward with your reference manager.

Box 1 | 1930s decision tool quality beat 1990s decision tool quantity in antibacterials

"Is it not peculiar that the first useful antibiotic, the sulphanilamide drug Prontosil, was discovered by Gerhard Domagk in the 1930s from a small screen of available dyes (probably no more than several hundred), whereas screens of the current libraries, which include ~10^7 compounds overall, have produced nothing at all?"[35].

Figure 3a suggests it is not peculiar, provided Domagk's decision tools had much higher predictive validity than those that were employed seven decades later. Using the conditions in Figure 3a, and sampling from the same population of compounds, the best molecule out of 200 chosen using a decision tool with a predictive validity of 0.8 is more likely to work in people than the best molecule out of ~10^7 from a decision tool with a predictive validity of 0.2.

Domagk tested drug candidates against *Streptococcus hemolyticus* in bacterial cultures and *in vivo* mouse models. In 1932, he found an azo dye (KL-695) that protected mice without *in vitro* activity (it is metabolized into its active form). Optimisation led to KL-730, sulfanilamide (Protonsil), which found clinical use in streptococcal infections[36]. Less successful discovery efforts in the late 1990s and early 2000s have been described by several authors[35,38–41]. By 1995, the industry had shifted to target-based discovery, often with high-throughput screening (HTS). The targets came from genes that were essential to the survival of a range of pathogens and that lacked close homologues in people[35,38,39]. GlaxoSmithKline (GSK) conducted 67 HTS campaigns, each with up to 500,000 compounds, and found 16 hits and 5 leads, but no broad-spectrum candidates that were worthy of clinical trials. Furthermore, *"GSK was not the only company that had difficulty finding antibacterial leads from HTS… > 125 antibacterial screens on 60 different antibacterial targets were run by 34 different companies… [but none] resulted in credible development candidates…"*[38].

So why did Domagk's mice outperform genomics and HTS more than six decades later? One possible explanation is chemistry. HTS collections circa 1995 were enriched for compounds that are unlikely to lead to good antibacterial drugs[35,38–40,254,255]. However, any chemical explanation must be held up against the limited chemistry available to Domagk circa 1930. A more plausible explanation is the increase in difficulty from the shift from narrow-spectrum sulfanilamide to broad-spectrum candidates. However, the drug industry had great success with broad-spectrum antibiotics from the 1940s to the 1960s, using decision tools that Domagk would have recognized.

We believe that the adoption of HTS and target-based discovery depressed both the congruence between the decision tool and the human disease state ("biological recapitulation" in Table 2) and the

congruence between testing protocols and the human disease state ("tests and endpoints" in Table 2) such that the results were largely irrelevant for human disease. The decision tools' "domains of validity" (Table 2) no longer included clinical utility in people. The main biological sources of de-correlation are probably bacterial efflux pumps and permeability barriers[35]. The compounds that scored well in HTS campaigns against isolated bacterial proteins were unlikely to accumulate in bacterial cells in sick people[254]. The de-correlation was probably compounded by a mismatch between the genes that were essential for survival in the bacteria in the genomics studies and the genetic and metabolic circuits that are important for bacterial survival in a mammalian host[37,38,40]. In contrast, drug candidates that appeared attractive in Domagk's infected mice were already likely to enter bacterial cells and act on the relevant machinery.

Since the failures of the first wave of target-based methods, there has been a return to screening against live bacterial cells but in ways that provide mechanistic insights very early in the process[37,40,255,256]. This may provide the best of both words[256]; whole cell activity to avoid compounds that fail to accumulate in live bacteria, plus defined mechanisms to reduce the effort wasted on non-specific toxins and for efficient optimisation of any hits. Furthermore, much of this can now be done at very high throughput[255,257], potentially offering both quality and quantity (see upper right quadrant of Figure 3a).

Box 2 | Decision tool innovation for hepatitis C virus infection spurred therapeutic innovation

Hepatitis C was recognized as a distinct viral liver infection in the early 1970s, with the causative virus isolated in 1989. However, the hepatitis C virus (HCV) was extremely difficult to grow in cell culture[43,44,258], which, in turn, made it impractical to develop useful decision tools for developing antiviral drugs. There were reports of HCV infection in cultured cells, but viral yields were so low and variable that it was often hard to distinguish between the RNA of the virus that was used for the infection and the RNA that resulted from the infection[42]. Transfection models, introducing isolated viral RNA into cells, also suffered from similar signal to noise problems that rendered them of little use[42]. Some HCV proteins could be expressed and investigated, but models using isolated proteins tend to be much less predictive of clinical antiviral activity. The dearth of cell-based decision tools proved a real R&D roadblock[42,43,258,259].

A decade after the cloning of HCV, after many false leads, and building on work in other RNA viruses[260–262], Lohmann et al.[263] reported the creation of HCV "replicons". Replicons comprise a truncated viral genome. They lack code for the structural proteins of the full virus but retain genes for the machinery of viral RNA replication and sometimes add non-viral reporter genes (for example, firefly luciferase). The reporter genes identify cells in which viral RNA replication is underway and signal the quantity of viral RNA replication.

Parallel selection of cell-culture adaptive replicons and replicon-friendly cell clones led to measurable, reliable, high titre, viral RNA replication in a range of cell lines. Replicons formed the basis of high-throughput phenotypic screens[45], efficacy assays for lead optimization[45] and antiviral drug resistance models. They were critical for the discovery of most, and the optimization and preclinical development of all, of the first wave of direct-acting antiviral HCV drug classes[42,45,258]: NS4/4A protease inhibitors (such as boceprevir and asunaprevir), NS5B polymerase inhibitors (such as sofosbuvir and dasabuvir) and NS5A inhibitors (such as ledipasvir and daclatasvir). These drug classes transformed treatment, with a shift from relatively intolerable and ineffective interferon-based regimens, to regimens that are both tolerable and rapidly curative[264].

In our terminology, replicons traded a decline in 'biological recapitulation' versus other cell culture methods (Table 2) — losing some native viral machinery and the full viral replication process — for huge gains in 'statistical and experimental hygiene' (Table 2). But, in contrast to isolated HCV proteins, replicons recapitulated enough of the relevant biology to support decision tools whose 'domains of validity' (Table 2) included clinical utility in man across several drug classes.

Box 3 | Translation failures in ischaemic stroke

Reviewing the field in 2007, Sena et al.[130] wrote: *"At least 883 candidate thrombolytic and neuroprotective drugs have been tested in animal models of stroke and shown some evidence of efficacy; 97 of these drugs have been tested in human ischaemic stroke. To date, there is unequivocal evidence for efficacy of only two drugs; aspirin and tPa [Alteplase]."*

This should be a surprise. After all, we know with a high degree of confidence what causes ischaemic strokes and can, in principle at least, replicate the cardiovascular consequences in animals.

There is now a large body of work, much related to what has become the CAMERADES collaboration, that bears on the translation failure[129,130,271,131,135,265–270]. First, animal studies often used "tests and endpoints" (Table 2) that failed to map onto clinical practice in humans. For example, most animal studies have taken a volumetric approach to lesions, while pivotal human trials focus on functional recovery[130,269]. There have been problematic dosing differences between animals and man[130] since drugs[269] that damp-down excitatory toxicity in oxygen-starved neurons often have cardiac and/or neurological toxicity[269]. Strokes strike rapidly and there were often vast differences in the timing of the pharmacological intervention with respect to the ischaemic injury in animal models compared to what could be achieved in clinical practice. For example, across 19 published animal studies of the drug candidate tirilazad[130], the median delay between the ischaemic event and the start of drug treatment was 10 minutes. In failed human trials of the same drug, the median delay was 5 hours. Of course, it is possible that the 'domains of validity' (Table 2) of some preclinical tirilazad models would extend to human stroke treatment in the unlikely event that people could be treated within 10 minutes of the onset of ischaemia. In fact, we now know that for the clot-buster, tPa, the window of therapeutic opportunity in well conducted mouse studies is within 3 hours of the ischaemic event[272]; similar to the ~4.5 hour window for treatment in stroke patients[273].

Second, like antimicrobials in the 1990s (Box 1), the field was dominated by test systems that were relatively convenient and inexpensive, in this case rodent-based models, but at the cost of questionable 'biological recapitulation' (Table 2)[269,270]. Glucose and oxygen metabolism and cerebral blood flow are three times higher, per unit volume of brain tissue, in rats than they are in people[269]. White matter is less than 10% of the rat brain but ~50% of the human brain[270]. Rodent studies focused on grey matter lesions, while patients in trials typically have substantial white matter involvement[270]. Rodents' grey matter is also different to humans' in terms of the neuron to glia ratio (1:2 in rodents, 1:10 in humans)[270], and ischaemia has different effects on neurons and glia. Then there are major species differences in cerebral vasculature[269]. Furthermore, some of the anaesthetics used in rodent stroke models appear themselves to be neuroprotective[267]. It now seems likely that non-diabetic

normotensive[267] rodents respond better to a range of neuroprotective drugs than do the bulk of elderly human stroke patients[269,270].

Third, ‘statistical and experimental hygiene’ (Table 1) was often poor; at least in published animal studies of ischaemic stroke[130,266–268]. This generates false positives in at least two ways[95]. First, it introduces random noise, some of which is positive[94,95] (for example, inadequate sample size[130,267]). Second, it introduces systematic bias via an unwarranted focus on the positive component of the random noise (for example, publication bias[266]) and/or by artificially inflating effect sizes (for example, via unblinded assessment of animal outcomes[267]).