# Overcoming psychological barriers to good discovery decisions


## Andrew T. Chadwick and Matthew D. Segall

Dr A T Chadwick,
Tessella plc
Bretby Business Park
Ashby Rd
Stanhope Bretby
Burton upon Trent
Staffs
DE15 0YZ, UK

Dr M D Segall,
Optibrium Ltd
7226 IQ Cambridge
Beach Drive
Cambridge
CB25 9TL, UK

Corresponding author: Chadwick, A T. (andrew.chadwick@tessella.com, fax 01283 559151)



This is a preprint of Drug Discovery Today, Volume 15, Numbers 13/14, July 2010.

**Abstract**
Better individual and team decision-making should enhance R&D performance. Reproducible biases affecting human decision making, known as cognitive biases, are well understood by psychologists. These threaten objectivity and balance and so are credible causes for continuing unpleasant surprises in Development, and high operating costs. For four of the most common and insidious cognitive biases, we consider the risks to R&D decision-making and contrast current practice with use of evidence-based medicine by healthcare practitioners. Feedback on problem solving performance in simulated environments could be one of the simplest ways to help teams improve their selection of compounds and effective screening sequences. Computational tools that encourage objective consideration of all of the available information may also contribute.

**Introduction**
Drug discovery leaders receive much conflicting advice on possible ways to improve productivity and restore past levels of return on investment [1, 2]. Competition from generics has strengthened and there is a lack of low-hanging fruit [3] in terms of validated targets for currently untreated diseases. Continuing investment in predictive science and translational medicine, outsourcing of shared services and formation of smaller, disease-specific units that bring researchers closer to clinicians, are all current trends, together with attempts at continuous improvement.

One common approach to improving productivity, reducing cycle time through restricting the volume of on-going work (work-in-progress), will demand particular precision in early attrition decisions. The Six Sigma movement, with its heritage in manufacturing, is making progress on time and cost saving for repeated discovery tasks [4]. However, standard recipes cannot be applied to decision-making within projects that face unique challenges and constraints [5]. Even so, senior management cannot afford to ignore the human dimension – are their teams making the best possible decisions given the information available to them, or that could be available given the right experiments?

Psychology research [6] has proved, again and again, that humans are inherently weak at making complex choices and plans, where there are a variety of sources of risk and also uncertainty about both the amount of each risk and the potential payoff. Even though efficient reduction of uncertainty is central to good research [7] (once past the initial creative stage), many practical researchers remain baffled or confused by probabilistic models and so shy away from formal decision analysis. Yet reliance on gut instinct tends to lead to consistent patterns of mistakes. Human nature means that we are all too quick to seize at something that looks initially promising and to run with it despite mounting negative evidence; we are over-ready to justify our own past decisions, seeking evidence that will support our past judgment rather than critique it; and we have short-term memories and attention spans that over-emphasize recent and newsworthy information.

What's worse, even when we know these 'thinking traps' [8] or ‘decision traps’ [9] in theory, we still continue to fall into them and can only learn to improve through practice [10]. Inexperienced bridge players tend to overbid on their first promising hand; for them, it is the consequence of letting down their partner, rather than textbook theory, which gives effective feedback. Why should discovery scientists be any different in their decision-making habits?

Given the long timescale of pharma R&D it is hard, if not impossible, to find out empirically and learn what makes for success and failure through personal experience alone. Discovery groups, therefore, need to define and encourage 'best practice' for project conduct in a way that captures wider company and industry experience. There is a need to make this as simple and accessible as possible, providing not hard rules but guidelines that can be flexed for circumstances, and can be applied within practical timescales and under management pressure to make rapid progress.

However, unfamiliar and sometimes counterintuitive concepts are involved. Success or failure can involve more luck than judgment (http://www.creatingtechnology.org/biomed/chance.htm) and yet sustained value creation relies on placing bets well on the options within each project, over a number of projects, most of which will fail along the way [11]. Furthermore, best practice will need to be tailored to individual research and disease areas according to the acceptable product profile, allowing for differing sources of risk within the relevant chemotypes. Many decisions on candidate progression also need to recognise that a target product profile, which is a view of future market conditions, is itself subject to error and uncertainty.

The move to project-specific screening choices may be a hard transition for organisations that have a fixed process culture: “Decision-making is managed simplistically by following pre-established, expected outcomes at so-called “go–no-go” decision points. It is black and white; the mentality is that there is no need to agonize over decisions. Avoiding the thinking process does not serve research well” [12].

Selecting and maintaining a flexible set of rules for choosing hits, leads and candidates is non-trivial given the increasing variety of available tests, including new high-throughput methods for early ADME and toxicology screening. The practice of clinical medicine faces similar choices - how much screening and early prevention to attempt, given that false positives can cause unnecessary anxiety, expense, and actual harm (e.g. unwarranted side-effects from prostate surgery in advancing age). In drug discovery, false alarms can mean throwing out perfectly good compounds in error, reducing the opportunities to find new therapies and sources of profit. Both in medical practice and discovery, each test also has a cost to be considered, either a variable cost or a fractional loading on fixed capacity.

What are the most promising approaches to help teams make decisions in a way that optimizes overall discovery performance? Drawing on our own experience in both pharmaceutical research and the application of evidence-based medicine, we consider and contrast some of the most relevant sources of cognitive biases and approaches to mitigating these. Each row in the following table summarizes sources of irrational decisions that have been studied and confirmed within experimental psychology [13]. For each source, we will consider evidence and ideas from drug discovery and from the practice of medicine, looking at ways of helping people work in a way that is truly ‘evidence-based’:

| Bias | Drug discovery implications | Medical implications |
|---|---|---|
| **Over-confidence**: reluctance and/or inability to 'prove a negative', and premature closure | Projects failed too late. Insufficiently wide search before choosing a lead series or candidate. | Diagnostic error and inappropriate course of treatment |
| **Poor calibration** of judgments on estimating and forecasting reliability. | Inappropriate weight given to results from high-throughput or *in silico* early screening methods, or rules of thumb, that may too often reject drugs that should have survived | Inadequate attention to the balance between the risks of inaction and action (e.g. use of biopsies) |
| '**Availability' bias**: over-attention to the vivid and recent, with neglect of prior information | Failure to apply and learn from the 'big picture' of industry project successes and failures | New clinicians are too prone to consider rare exotic diseases as the cause for observed symptoms |
| **Excess focus on certainty** when considering whether to accept one or more sources of residual risk | Inefficient use of resources when screening across multiple risk factors or possible indications | Difficulty in agreement and use of clinical guidelines; problems in reassuring patients. |

**Over-optimism and premature closure can lead to insufficiently wide search**

*"A person is never happy except at the price of some ignorance*" (Anatole France)

The history of science and medicine is full of wrong ideas that prevailed for many years despite mounting evidence to the contrary: phlogiston, the four humours, spontaneous generation of life, or inheritance of acquired traits. These are examples of 'confirmation bias', which means that "we tend to subconsciously decide what to do before figuring out why we want to do it" [8] and seek evidence that tends to confirm rather than refute our initial judgment (see Box 1 for an illustration). In medicine, such 'bad science' [14] can cost many lives. Therefore, major institutions and professions have procedures and rules, notably peer review, which seek to protect against the pernicious effects of excessive self-confidence (setting aside, here, the issue of blatant fraud.)

**Box 1.** *What's the rule?*

The sequence 2, 4, 6 is an example that obeys a general rule… but what exactly is this? To probe what the rule might be, you can specify other sequences of three numbers and ask if they obey the unknown rule. When you're confident of your hypothesis, you can announce what you think it is.

The question is… what sequences would you choose?

When this question was posed by Peter Cathcart Wason [42] in a famous experiment in 1960 he found that the subjects often announced complex rules after testing their hypothesis with sequences that obeyed their hypothetical rule rather than testing their rule with a sequence that violated it.

In this example, the general rule was "any ascending sequence"!

In our direct experience, discovery scientists admit that false optimism helps keep them functioning despite the recognized reality that most of their projects fail. This seems an essential trait of scientific heroes of the past [12] yet, paradoxically, may count as a cognitive error in a business setting [15]:

"Given the high cost of mistakes, it might appear obvious that a rational organization should want to base its decisions on unbiased odds, rather than on predictions painted in shades of rose. However … optimistic self-delusion is a diagnostic indication of mental health and well-being ... The benefits of unrealistic optimism in increasing persistence in the face of difficulty have been documented …[16]. The observation that realism can be pathological and self-defeating raises troubling questions for the management of information and risk in organizations. Surely, no one would want to be governed entirely by wishful fantasies, but is there a point at which truth becomes destructive and doubt self-fulfilling?"

Major discovery decisions cover targets, libraries, screening plans, lead series and candidate selection. Each of these has potential thinking traps, notably post-hoc rationalization and the tendency to cling too long to an idea, in part owing to considering 'sunk costs' rather than just future costs.

Choosing a target that will correspond to effectiveness in man is still the hardest part of drug discovery because of the sheer complexity of physiology and pharmacology and the partial nature of scientific understanding of pathways and genomics. This leads to an inevitable element of chance in finding relevant targets; new science is a source of unknown unknowns (black swans).

Therefore, when choosing compounds, where there are greater opportunities for learning from past experience, it is doubly important to avoid failure from more predictable causes, to make a sufficiently complete search amongst alternatives, and to use selection rules that incorporate the most recent information on all the targets to which binding is observed.

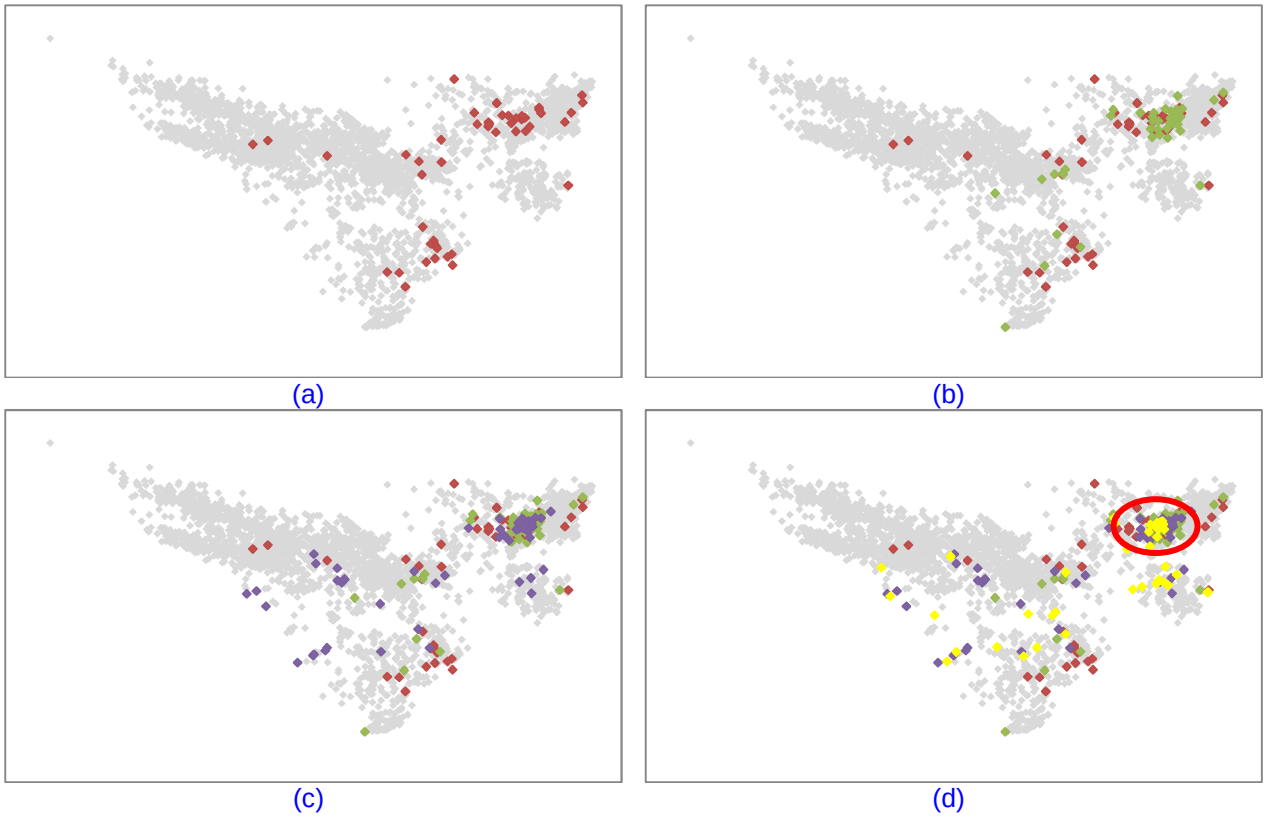



**Figure 1** These 'chemical space' maps illustrate the distribution of compounds selected for progression to secondary studies in a project. Each point represents a compound and the proximity of points indicates their similarity of chemical structure. The grey points illustrate the full diversity of the compounds screened in this project and the coloured points show those for further study in chronological order in the order red, green, blue and yellow in plots (a) through (d) in groups of approximately 50. The project had difficulty finding a compound with appropriate properties and, in light of this, these maps suggest that too much weight was given to the region circled in plot (d) rather than searching more widely for a satisfactory compound.

There is evidence within lead discovery and optimization that people trust their judgment too much and focus early on what looks promising to them, rather than spreading their search widely enough (see Figure 1 for an example). In this project, targeting an orally bioavailable compound for a central nervous system (CNS) target, early pharmacokinetic (PK) data showed that compounds in the region circled in Figure 1 (d) could have either good oral bioavailability *or* good penetration into the CNS. This suggested to the team that similar compounds may exhibit *both* desirable properties simultaneously. Therefore the project returned repeatedly to the same chemistry in the expectation that an optimal compound could be found. Only after progressing almost 200 compounds for detailed *in vitro* studies and approximately 50 compounds for *in vivo* PK studies was an alternative strategy pursued in earnest. More details on this example can be found in Segall *et al*. [17].

In clinical practice, the pitfall of over-confidence is even more directly a matter of life and death. An analysis of diagnostic error in internal medicine [18] covering 90 injuries, including 33 deaths, showed that cognitive factors contributed to diagnostic error in 74% of cases: "Premature closure, i.e. the failure to continue considering reasonable alternatives after an initial diagnosis was reached, was the single most common cause."

One important initiative in the UK to encourage evidence-based practice is the Map of Medicine (see http://www.mapofmedicine.com ™Hearst Corporation) Joining up the efforts of clinicians in primary and secondary care, this "provides a visualisation of the

ideal, evidence-based patient journey for common and important conditions ... The Map is a web-based tool that can help drive clinical consensus to improve quality and safety in any healthcare organisation". The clinical pathways branch according to findings and assessments on the individual patient. They include triggers for urgent action (red flags) and criteria to be applied in the choice between diagnostic testing, which may involve referral to specialists, or a 'wait and see' approach. Individual clinicians can provide feedback on any pathway to an expert panel and curation of pathways is now taking place at the local level to reflect availability of resources such as expert physiotherapy that can avoid the need for referral to hospitals (e.g. for joint injections). Many of these pathways are also now directly available to patients (e.g. http://healthguides.mapofmedicine.com/choices/map/index.html ) to help reassure them about their treatment path.

An equivalent system for pharmaceutical R&D would consist of curated, evidence-based screening plans: a library of screening pathway options with criteria and interactive support for individual projects to make appropriate modifications to meet their needs. A hard challenge for today's project teams and managers is how – given what may be limited method calibration data relevant to their new project and judgments based on extrapolation from previous experience – to work out the right screening pathway to choose, that balances the risks and consequences of false positives and of false negatives and uses overall resources (including time) wisely.

**Poor calibration of the quality of predictions can lead to faulty balance between the risks of action and inaction**

*'Prediction is very difficult, especially if it's about the future'* (Niels Bohr)

A particular kind of over-confidence applies to estimates about the future. Individuals' calibration of the quality of their own predictions is notably inaccurate [19]. People asked to make an estimate or prediction – say, the closing value of a stock market index on the following day – and then asked to give a range that, 95% of the time, will include the correct answer, consistently provide too small a range. Therefore, in drug discovery, decision-makers considering how to eliminate compounds with undesirable properties are likely to underestimate the importance of understanding the trade-off between false negatives and false positives [20]. This means they will tend to incur excessive costs of late failures or will lose opportunities to develop valuable products.

Over-confidence in the power of one's methods can happen within Development as well as in Discovery. "The constant accrual of new data requires an ongoing assessment of the benefit/risk profile of a compound to be made and for predictions … to be adjusted … These decisions are difficult to calibrate and can have profound sequelae" [3]. For example, the sustained and expensive lack of success in clinical development within certain therapeutic areas such as stroke mitigation (other than thrombolysis) [21] might be due to one or more of:

- unrealistic levels of optimism,
- too much faith in animal models [22] (a human lesion can be larger than a rat brain, or even a whole rat)
- publication bias in favour of positive results for animal trials.[23]
- a lack of basic understanding of the causal factors at work [24]
- inadequate statistical power in the experiments and trials [25]

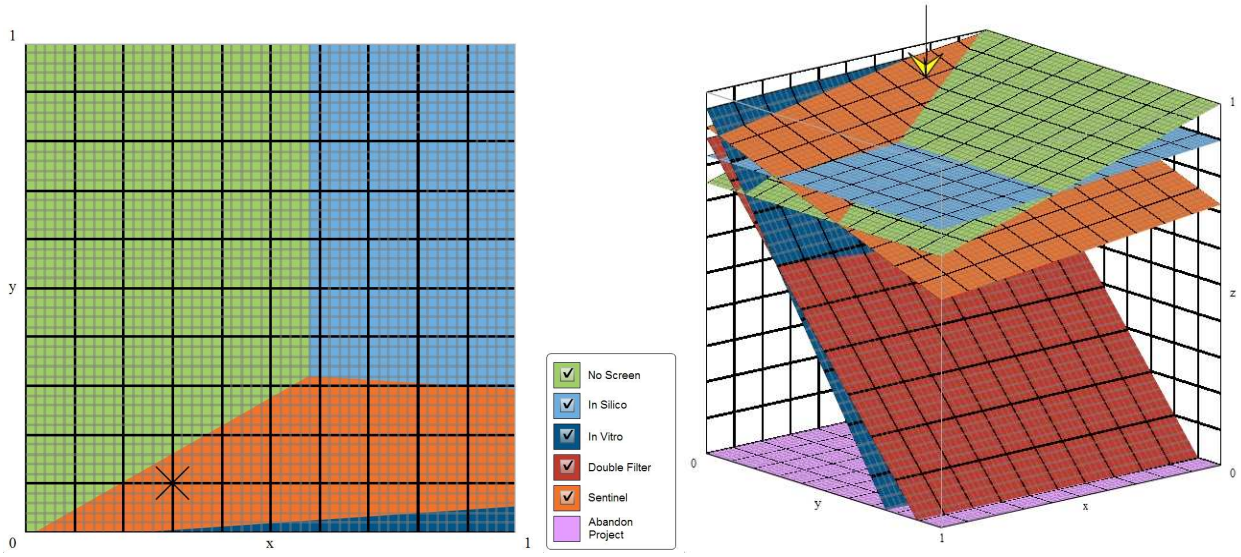
Cognitive biases are deeply ingrained and hard to overcome in medical practice [26], as well as within research. Individual awareness of calibration bias is not enough to sidestep it. More effective measures could include team review, computer-assisted presentation of relevant information [27], or practice in a realistic but safe environment with accelerated feedback using a reference class of problems that are similar to the ones the team is likely to face. For example, in the final stages of radiology training and beyond, practitioners are encouraged to calibrate their judgment on examples of test cases. This helps them to understand how good they are at making judgments based on partial data, so they can accordingly base their advice to discharge or progress the patient to invasive tests or treatment. For breast cancer radiographic screening in the UK there is a 'round robin' exchange of blinded test cases, selected to cover the full range of required diagnostic discrimination, with feedback on achieved performance [28]: "Tracking and reporting critical outcome measures, such as sensitivity, specificity, size and stage of tumors detected, interval cancer rates, and time to recall and diagnosis, have been used in many countries to improve screening performance."

Many examples, with feedback, are needed to improve the ability to reflect accurately on how well they know what they know. Accuracy tends to improve through repetitive forecasting with short-term feedback – such as weather forecasting – relative to single, one-off forecasts [29]. This is a particular challenge for pharmaceutical R&D, given that many years may elapse between choice of target and validation of that target within Phase 2 trials.

This makes it essential to transfer experience from one project that can look back on its mistakes to another that still has the chance to learn. This learning often cannot be about project specifics, given the variety between projects. Instead, it is ‘metaknowledge’: reasoning ability about how much we know and still need to know (a form of wisdom, according to Plato). Such knowledge will be under broad headings − such as sources and collective estimates of risk, practical reliability of predictive methods, and reasons why predictions may be confounded − rather than knowledge about specific targets. Hence, practical support for better R&D decision-making has to enhance the scientific method of sceptical enquiry rather than attempt in any way to replace or automate it.

R&D decision-making and planning skills can be built up from exploring problems that capture key elements of the structure of the real-world challenges; for example the need to consider impact of possible errors in predictions. Many scientists prefer to learn by doing rather than from theory. If feedback is sufficiently rapid, this will also be more effective in overcoming built-in biases. Training simulations which capture essentials of the cognitive challenge, and provide feedback on performance are sometimes referred to as ‘microworlds’. These have been used with success for training medical teams, especially for emergency response, for which experience is not routinely available and the real world does not give time to reflect and mentally rehearse courses of action before that action is needed [30]. Figure 2 illustrates such a microworld intended to support reasoning and judgment within drug discovery teams deciding on a screening strategy. The interactive version can be accessed without charge at http://www.tessella.com/screening-strategy-explorer. Considering just one source of hazard within this example, quantitative analysis of potential screening cascades can identify the most effective option. Formally, this maximizes the expected return on pipeline opportunities and on resource use within screening (which must also include

investment of elapsed time). Estimating screening cost, downstream failure cost and value of a safe compound often presents an obstacle for teams early in the R&D process. Fortunately, the cost variables need be considered only in terms of ratios to potential value. Furthermore, all feasible combinations of these ratios are viewed at the same time, so that teams can easily see for themselves whether these quantities statistically significantly impact the best decision and then make judgments, or obtain better estimates, accordingly.




(a) (b)

**Figure 2.** Presentation of the best screening strategies against a single hazard. The strategies employ two independent methods that have pass/fail outcomes and can be combined in sequence either as a double filter (progressing compounds that pass both methods) or using the sentinel approach (which terminates only those compounds that fail both methods). The plots reveal an analysis of the possible screening strategies: **(a)** sensitivity of best strategy to the two main cost parameters defined relative to the value of a successful compound, and **(b)** visualisation of the impact on pipeline value of the different strategies. Where two tests are used in sequence, the first is *in silico* and the second *in vitro*. The *in silico* cost per compound is assumed to be negligible.

*X axis*: the cost of a downstream failure due to this hazard, relative to the net value of a safe compound that has reached the same point in development
*Y axis*: the cost per compound of the *in vitro* test relative to the net value of a safe compound exiting from screening
*Z axis*: the value of each screening strategy relative to the ideal strategy which would filter out all truly unsafe compounds at zero cost. This ideal strategy value is reduced by the cost of compound screening, the pipeline impacts of false alerts (lost value potential), and missed alerts (incurring increased downstream cost of avoidable failure)

In this example, the model parameters are

- Prevalence of the toxicity problem in the population of unscreened compounds: 20%
- Sensitivity of the tests: 70% for *in silico* and 90% for *in vitro*
- Specificity: 90% for *in silico* and 98% for *in vitro*

The broad orange region arrowed, at realistically low values of Y, indicates that the 'sentinel' combination of *in silico* and *in vitro* screens will usually be favoured over either alone at these values of risk and method performance, unless the *in vitro* testing cost is particularly low. The 'double filter' strategy (red plane in b) never outperforms the 'sentinel' strategy, at any combination of cost or value, for this (or lower levels of) risk. The strategy of using no safety testing (green plane) ahead of the *in vivo* (regulatory) safety assessment appears a realistic one for relatively high-value projects (where X tends to be low), due to the impact of false alerts from both the available *in silico* and *in vitro* methods. This 'no screen' strategy demands an acceptable opportunity cost of late failure, for example through use of multiple parallel candidates.

From our experience so far, using this simulation and in previous assessment of *in silico* toxicity methods, unaided teams are liable (at typically low underlying probabilities of a hazard) to miss a robust screening option, the 'sentinel approach'. This combination, wherein failures on both the *in silico* and the *in vitro* methods are needed to reject a compound or series, throws few false alerts, and can therefore take advantage of a pass hurdle (cut-off) which favours sensitivity over specificity (see for example http://www.aapspharmaceutica.com/meetings/files/36/Kreatsoulas.ppt#302,14,Performance Assessment: Has DEREK been Improved?).

**Vivid and recent events can unduly dominate thinking about relative risk**

*'Those who cannot learn from history are doomed to repeat it'* (George Santayana)

Individuals are biased towards recent, vivid experience and tend to ignore relevant information on long-run chances of a problem. As an illustration of this, it is believed that more people died on the roads after 9/11, as a result of increased road traffic caused by avoiding airline travel, than in the airplanes deliberately crashed. The many small tragedies are less vivid and available to the individual decision-maker than the one large and vivid event, distorting assessments of relative risk.

There is an especial danger that pharmaceutical research, always looking for better ways of working, will pay too much attention to recent information rather than the sum of all relevant data. This cognitive 'availability bias' is sometimes termed 'neglect of the prior.' The prior is the underlying probability of occurrence of an event in the absence of new evidence.

Team members can also be over-influenced in their own beliefs by the opinions of outspoken or powerful individuals, hence: "It is more important than ever that a leader's ability to make decisions be based on an understanding of probability with a capacity to recalibrate one's perspective in the light of novel information − so-called Bayesian thinking [3]"

People often do not factor in the track record of reliability of a prediction or diagnosis when acting on its conclusion; there is a tendency to put too much weight on this specific, recent information and not enough on prior information such as past outcomes for similar compounds or patients. If prediction reliability is not well known then there is the additional need to avoid the trap of calibration bias, which will tend to over-estimate this reliability and so further discount the relevance of the prior.

**Box 2.** *How well does this test conserve your compound options?*

You have purchased a series of compounds, within which you expect 1% have a particular kind of toxicity. You apply a screening method to all the compounds that is 90% reliable (both 90% sensitive and 90% specific: this means that if a compound is genuinely toxic there is a 90% probability of detecting this, and if it is not toxic, there is a 90% probability that the test will report the compound as being safe).

What percentage of the compounds that fail the screening genuinely have the toxicity?

a) About 1%
b) About 2%
c) About 10%
d) About 50%
e) About 90%

In medicinal chemistry, one danger from availability bias is excessive attrition to the pipeline through placing too much reliance on faint but uncertain signals of, for example, toxicity, as illustrated by the question in Box 2. (The answer appears in ref [31]). Gigerenzer [32] and others have shown that probability ratios expressed as natural frequencies e.g. 2 out of 1000, tend to lead to better judgments than the percentages provided in this example.)

Unless there is diversity to spare, these hazard early warning signals need to be handled with care and, where they would restrict choices on the way forward, should be trusted as the sole basis for decisions only for very reliable predictions, so long as it is ethical to do so (i.e. where a later regulatory animal test can be relied on to protect subsequent human volunteers).

This dilemma faced by discovery scientists – whether to abandon a line of research because of some predicted risk – is also common in medicine, especially in routine non-invasive screening. For example, when questioned on the diagnostic weight of positive AIDS test results or breast cancer signs in younger women, a majority of medical students considerably overestimate the fraction of alerts that are genuine through neglect of the information about frequency in the overall population. Given some level of a signal indicating a potential problem, the question is, how aggressively to follow up. A false positive means worry to the patient and the pain and accompanying (small) risk of a follow-up biopsy or other invasive test; a false negative could delay the definite confirmation and possible cure of a serious or fatal disease.

In medicine, it is well known that recently qualified practitioners have a tendency to over-diagnose exotic but rare conditions which they have heard about during training - as if every day was an episode of the television medical drama 'House'. Where a symptom could be of a rare tropical fever, or an unusual presentation of flu, more experienced doctors will opt for the diagnosis of flu.

The correct decision stance depends in part on the consequences of a mistaken assessment. It also depends on good use of information on prevalence of the problem in a suitable reference population (similar patients, or similar projects). Choosing the reference population for a drug discovery project is non-trivial, as the larger the sample size the less relevant some of the past examples may be to the biological, chemical or developability challenges within the current project.

Decision-making on choice of compounds and on choice of further screening therefore has to be based on a mix of evidence and judgment. Teams can be helped in their reasoning and synthesis of the evidence through interactive data mining and through 'what-if' analysis of pipeline filtering options using cascade simulations or more rapid, but often less intuitive, decision tree approaches. Prior probabilities will be an important input for such 'what-if' simulation systems. Estimated probabilities for different hazards, or for predictive method error, need to be taken not only from industry benchmarking (which reflects the outcomes of current screening practices) but also from a proportion of in-house effort dedicated to breaking the normal rules (within ethical limits), following up seemingly unpromising options, and helping to calibrate both risk and the reliability of the current screening strategy. Such exploration efforts will also be accompanied by a small percentage of pleasant surprises – drugs that turned out safe and active despite initial appearances.

Robert deWitte [33] has compared drug discovery to the Tour de France; the overall winner need not have won in any one event, but is an 'all-rounder' over both the mountains and the sprint stages. There is a need for a balanced judgment, taking into account the several desirable characteristics of a drug, and the degree of certainty about quality on each of these criteria from the evidence accumulated to date. This is a form of 'multi-attribute' decision making.

**An excessive focus on certainty sometimes warps the distribution of effort**

*"I prefer the errors of enthusiasm to the indifference of wisdom"* (Anatole France)

One of the less-recognized, but consistent, errors in multi-attribute human decision-making is an excessive focus on certainty [34]. The psychologists Daniel Kahneman and Amos Tversky developed an empirical model of how people actually take decisions under risk, termed 'Prospect theory'. In this model, people are more sensitive to changes in probabilities around 0 and 1 than, logically, they should be. We seem at the same time to be both risk-averse and risk-seeking; we might buy both insurance (to be sure of avoiding a loss, when perhaps we could afford to absorb it) and also a lottery ticket (overvaluing a low probability of gain). Both might involve overestimating the importance of probabilities that are close to zero.

There have been many academic and practical studies of multi-attribute decision-making. One notable conclusion has been that human decision-makers or 'judges' are inconsistent even in applying the rules they would describe if questioned on the basis for their decisions [35]: "the overwhelming conclusion, including studies of clinical judgment, was that the linear model of the judge's behaviour outperformed the judge."

A rational approach to screening systematically improves the odds through a cascade of risk assessments of increasing precision, taking first in sequence [36] the lower-cost and faster tests that are able to fail more bad compounds, provided the tests are sufficiently reliable. Many factors can be considered in parallel if the methods have a low enough cost and provide complementary information. This tends to spread the effort for risk reduction across many possible causes of failure, overcoming the bias towards seeking certainty on just the factor that leads to the most frequent late failures. An improvement in chance of success from 50% to 60% on one risk factor is worth just as much as an

improvement from 90% to 100% on another of equal impact, in this model of research performance.

One method of applying an efficient mix of methods is the probabilistic scoring approach employed by the StarDrop™ software platform to guide compound selection decisions in drug discovery. A probabilistic score indicates the likelihood of success of a compound against a set of property criteria, given the available property data for that compound and taking into account the underlying uncertainty in the data. The criteria might have different degrees of importance because, in practice, it might be appropriate to make a trade-off between properties if an ideal compound is not available (see Figure 3 for an illustration). Furthermore, uncertainties in the overall score are calculated and can be used to establish when one molecule can be confidently chosen over another [37] .

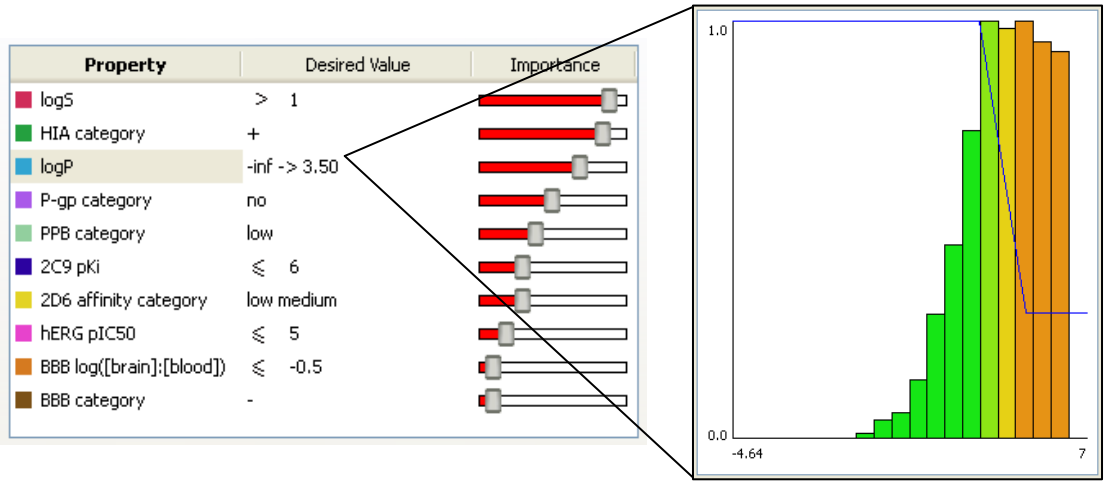
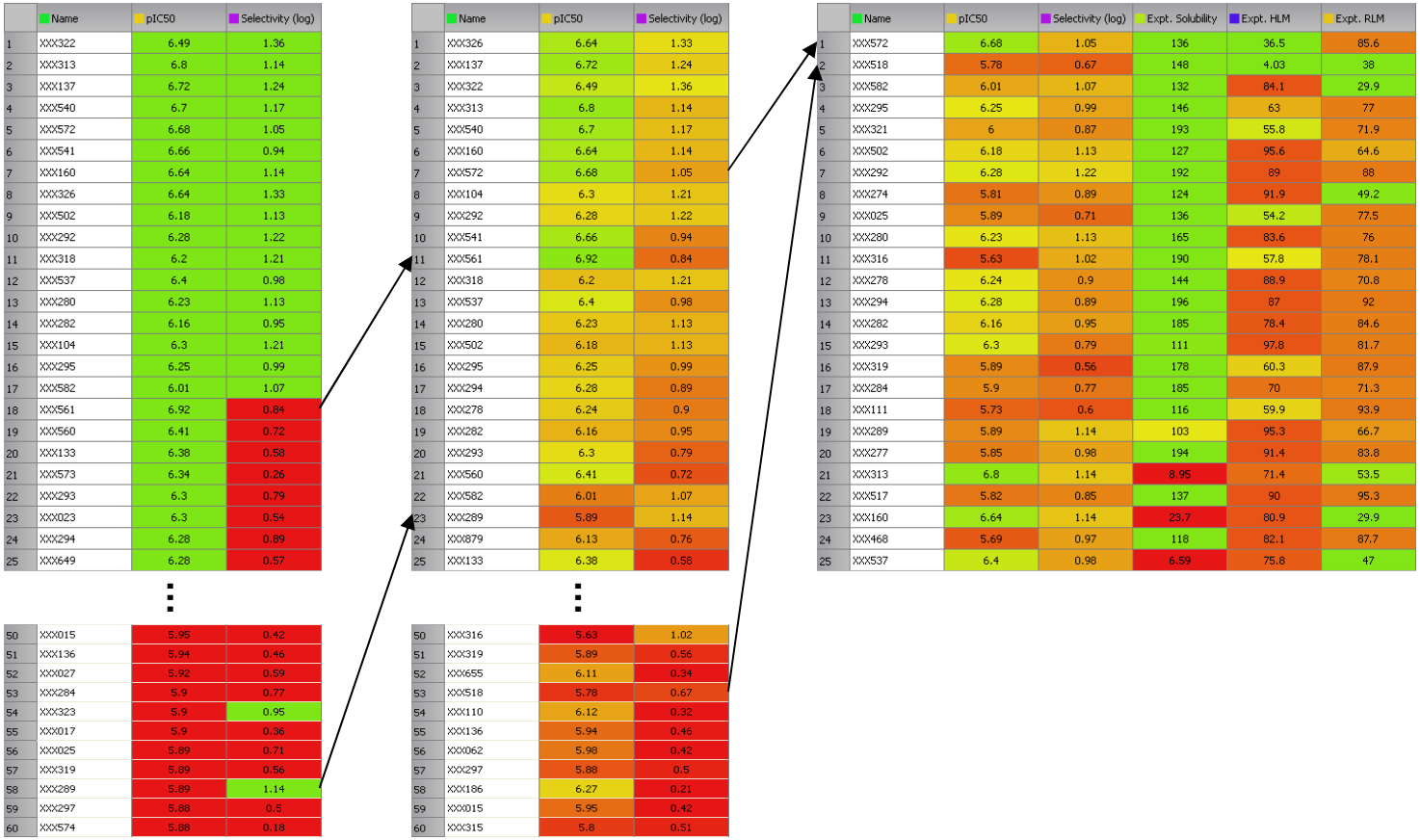



**Figure 3.** Example scoring profile, showing the ideal criteria and their relative importances (red bar). Furthermore, trade-offs can be defined that are more subtle than simple pass/fail criteria, because a scoring profile could contain more complex functions for each property representing a range of acceptability over the property value range. A gradual decline in quality relative to logP is shown, whereby a gradient has been specified between logP values of 3.5 and 5.0.

Here too the benefit of learning from experience is important, both in understanding the uncertainties inherent in different data sources and also, for each property, the relevance of a negative indication to clinical and commercial outcomes. The experience on predictive validity is transferable between projects; reasoning about outcomes will also require judgment on a project-by-project basis. Therefore, as discussed above, a proportion of effort should be dedicated to calibrate the risks due to each factor. A further advantage of this approach is that, if evidence arises that indicates that these assessments of risk should be revised; it is easy to rapidly assess the sensitivity of the choice of compound to any changes [38].

The advantage of explicitly considering a broader range of properties is illustrated in Figure 4(c), where compounds from a project have been scored for a balance of potency, selectivity and ADME properties. Comparing this prioritisation with that resulting from a narrow focus on potency and selectivity, as illustrated in Figures 4(a) and 4(b), shows notable differences in the compounds that would be selected. In particular, the second compound in the list, XXX518, which had previously been rejected, subsequently demonstrated a significantly improved *in vivo* profile over those compounds selected on the basis of potency and selectivity alone.



| | Name | pIC50 | Selectivity (log) |
|---|---|---|---|
| 1 | XXX322 | 6.49 | 1.36 |
| 2 | XXX313 | 6.8 | 1.14 |
| 3 | XXX137 | 6.72 | 1.24 |
| 4 | XXX540 | 6.7 | 1.17 |
| 5 | XXX572 | 6.68 | 1.05 |
| 6 | XXX541 | 6.66 | 0.94 |
| 7 | XXX160 | 6.64 | 1.14 |
| 8 | XXX326 | 6.64 | 1.33 |
| 9 | XXX502 | 6.18 | 1.13 |
| 10 | XXX292 | 6.28 | 1.22 |
| 11 | XXX318 | 6.2 | 1.21 |
| 12 | XXX537 | 6.4 | 0.98 |
| 13 | XXX280 | 6.23 | 1.13 |
| 14 | XXX282 | 6.16 | 0.95 |
| 15 | XXX104 | 6.3 | 1.21 |
| 16 | XXX295 | 6.25 | 0.99 |
| 17 | XXX582 | 6.01 | 1.07 |
| 18 | XXX561 | 6.92 | 0.84 |
| 19 | XXX560 | 6.41 | 0.72 |
| 20 | XXX133 | 6.38 | 0.58 |
| 21 | XXX573 | 6.34 | 0.26 |
| 22 | XXX293 | 6.3 | 0.79 |
| 23 | XXX023 | 6.3 | 0.54 |
| 24 | XXX294 | 6.28 | 0.89 |
| 25 | XXX649 | 6.28 | 0.57 |
| 50 | XXX015 | 5.95 | 0.42 |
| 51 | XXX136 | 5.94 | 0.46 |
| 52 | XXX027 | 5.92 | 0.59 |
| 53 | XXX284 | 5.9 | 0.77 |
| 54 | XXX323 | 5.9 | 0.95 |
| 55 | XXX017 | 5.9 | 0.36 |
| 56 | XXX025 | 5.89 | 0.71 |
| 57 | XXX319 | 5.89 | 0.56 |
| 58 | XXX289 | 5.89 | 1.14 |
| 59 | XXX297 | 5.88 | 0.5 |
| 60 | XXX574 | 5.88 | 0.18 |

(a)

| | Name | pIC50 | Selectivity (log) |
|---|---|---|---|
| 1 | XXX326 | 6.64 | 1.33 |
| 2 | XXX137 | 6.72 | 1.24 |
| 3 | XXX322 | 6.49 | 1.36 |
| 4 | XXX313 | 6.8 | 1.14 |
| 5 | XXX540 | 6.7 | 1.17 |
| 6 | XXX160 | 6.64 | 1.14 |
| 7 | XXX572 | 6.68 | 1.05 |
| 8 | XXX104 | 6.3 | 1.21 |
| 9 | XXX292 | 6.28 | 1.22 |
| 10 | XXX541 | 6.66 | 0.94 |
| 11 | XXX561 | 6.92 | 0.84 |
| 12 | XXX318 | 6.2 | 1.21 |
| 13 | XXX537 | 6.4 | 0.98 |
| 14 | XXX280 | 6.23 | 1.13 |
| 15 | XXX502 | 6.18 | 1.13 |
| 16 | XXX295 | 6.25 | 0.99 |
| 17 | XXX294 | 6.28 | 0.89 |
| 18 | XXX278 | 6.24 | 0.9 |
| 19 | XXX282 | 6.16 | 0.95 |
| 20 | XXX293 | 6.3 | 0.79 |
| 21 | XXX560 | 6.41 | 0.72 |
| 22 | XXX582 | 6.01 | 1.07 |
| 23 | XXX289 | 5.89 | 1.14 |
| 24 | XXX879 | 6.13 | 0.76 |
| 25 | XXX133 | 6.38 | 0.58 |
| 50 | XXX316 | 5.63 | 1.02 |
| 51 | XXX319 | 5.89 | 0.56 |
| 52 | XXX655 | 6.11 | 0.34 |
| 53 | XXX518 | 5.78 | 0.67 |
| 54 | XXX110 | 6.12 | 0.32 |
| 55 | XXX136 | 5.94 | 0.46 |
| 56 | XXX062 | 5.98 | 0.42 |
| 57 | XXX297 | 5.88 | 0.5 |
| 58 | XXX186 | 6.27 | 0.21 |
| 59 | XXX015 | 5.95 | 0.42 |
| 60 | XXX315 | 5.8 | 0.51 |

(b)

| | Name | pIC50 | Selectivity (log) | Expt. Solubility | Expt. HLM | Expt. RLM |
|---|---|---|---|---|---|---|
| 1 | XXX572 | 6.68 | 1.05 | 136 | 36.5 | 85.6 |
| 2 | XXX518 | 5.78 | 0.67 | 148 | 4.03 | 38 |
| 3 | XXX582 | 6.01 | 1.07 | 132 | 84.1 | 29.9 |
| 4 | XXX295 | 6.25 | 0.99 | 146 | 63 | 77 |
| 5 | XXX321 | 6 | 0.87 | 193 | 55.8 | 71.9 |
| 6 | XXX502 | 6.18 | 1.13 | 127 | 95.6 | 64.6 |
| 7 | XXX292 | 6.28 | 1.22 | 192 | 89 | 88 |
| 8 | XXX274 | 5.81 | 0.89 | 124 | 91.9 | 49.2 |
| 9 | XXX025 | 5.89 | 0.71 | 136 | 54.2 | 77.5 |
| 10 | XXX280 | 6.23 | 1.13 | 165 | 83.6 | 76 |
| 11 | XXX316 | 5.63 | 1.02 | 190 | 57.8 | 78.1 |
| 12 | XXX278 | 6.24 | 0.9 | 144 | 88.9 | 70.8 |
| 13 | XXX294 | 6.28 | 0.89 | 196 | 87 | 92 |
| 14 | XXX282 | 6.16 | 0.95 | 185 | 78.4 | 84.6 |
| 15 | XXX293 | 6.3 | 0.79 | 111 | 97.8 | 81.7 |
| 16 | XXX319 | 5.89 | 0.56 | 178 | 60.3 | 87.9 |
| 17 | XXX284 | 5.9 | 0.77 | 185 | 70 | 71.3 |
| 18 | XXX111 | 5.73 | 0.6 | 116 | 59.9 | 93.9 |
| 19 | XXX289 | 5.89 | 1.14 | 103 | 95.3 | 66.7 |
| 20 | XXX277 | 5.85 | 0.98 | 194 | 91.4 | 83.8 |
| 21 | XXX313 | 6.8 | 1.14 | 8.95 | 71.4 | 53.5 |
| 22 | XXX517 | 5.82 | 0.85 | 137 | 90 | 95.3 |
| 23 | XXX160 | 6.64 | 1.14 | 23.7 | 80.9 | 29.9 |
| 24 | XXX468 | 5.69 | 0.97 | 118 | 82.1 | 87.7 |
| 25 | XXX537 | 6.4 | 0.98 | 6.59 | 75.8 | 47 |

(c)

**Figure 4**. Tables illustrating three different views of the data for a single project.

- **(a)** Cells are coloured green if the compound 'passes' the required threshold for a property and red if it 'fails' according to the experimental data, but ignoring the uncertainty in the data. The pass thresholds applied are for potency (pIC50>6) and selectivity [log(selectivity)>0.9, equivalent to a factor of 8].
- **(b)** In this case, the uncertainties in the data are taken into account and the cells are coloured according to the probability that the compound passes the threshold for each property, from green (100%) to red (0%). The uncertainties in this case were estimated to be 0.5 log units for potency and 0.7 log units for selectivity. The compounds are ordered according to the probability that they pass the thresholds for both properties. When considered in this way, the ranking of some compounds changes significantly, as shown by the arrows for illustrative examples.
- **(c)** This table illustrates the effect of considering a broader view of the available data. In this case, compounds are ranked according to potency and selectivity, as in (a) and (b), but also solubility (threshold > 100µM) and stability in human and rat liver microsomes (threshold < 60%). The relative importance of these properties is also considered using the probabilistic scoring algorithm in StarDrop. When taking all of these properties into account, the effect on the selection of compounds is dramatic.

Prioritizing or weighting scoring criteria requires some common measure of impact: a metric to compare the factors being scored**.** Increasingly, health technologies are being chosen according to the value of information that they can provide [39]. Planning of experiments or the sequence of experiments [40] has an exact analogy with derivation of Clinical Practice Guidelines (CPG's). Gillian Sanders at Stanford has published powerful ways of deriving CPG's via decision trees and systematic capture of assumptions and evidence [41]. These ideas are catching on in the clinical community, and we believe that drug researchers will follow.

## Conclusion

As governments, insurers and health management organisations converge on outcome measures, such as a 'quality adjusted life year', we believe that the goals of pharmaceutical researchers, always broadly aligned with saving of life and suffering, will focus even more on medical benefit and so will become sharper and more useful as ultimate criteria for decision-making. Given these sharper goals, which are more scientifically accessible than a distrusted forecast of market potential or profit, there remains the challenge of making decisions that reach those goals as often as possible within the available resources.

Medicines have often been scarce in the past and even in the modern developed world are subject to some economic or policy-based principles of rationing. Discovery research has for long seemed to be in a world of plenty, but that world has changed. Learning to do more with the same, or with less, needs to be accelerated.

Only people can make good decisions. Helping people to reason better, to take on board the experiences of the past and to consider a wider range of options, requires special training and new approaches to analysing and presenting potential choices. The medical field is increasingly using a mix of new presentations of information (maps showing the choices of evidence-based paths) and simulations that provide rapid feedback on the actions that tend to succeed, or tend to fail, in the long run despite the confusing elements of chance. We think that pharmaceutical R&D needs to follow the same direction.

## References


1. Munos, B. (2009) Lessons from 60 years of pharmaceutical innovation. *Nature Rev. Drug Discov.* 8, 959-968
2. Paul, S. *et al* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Rev. Drug Discov.* 9, 203-214
3. Anon (2007) Why pharma must go Hollywood. *The Scientist* 21(2),42
4. Andersson, S. *et al* (2009) Making medicinal chemistry more effective – application of Lean Sigma to improve processes, speed and quality. *Drug Discov. Today* 14, 598-604
5. Tushman, M.L. and Benner, M. (2009) The productivity dilemma revisited: inherent conflicts between process management and exploration. In Adler, P.S *et a*l Perspectives on the Productivity Dilemma, *J. Ops Management* 27, 99-113
6. Baron, J. (2000) *Thinking and Deciding* (3rd edn), Cambridge University Press (paperback)
7. Peck, R.W. (2007) Driving earlier clinical attrition: if you want to find the needle, burn down the haystack. *Drug Discov. Today* 12, 289-294
8. Hammond J.S., Keeney, R.L. and Raiffa, H. (2006) The hidden traps in decision making. *Harvard Business Review* Jan, 118-126
9. Russo, J.E. and Schoemaker, P.J.H. (2002) *Winning Decisions: getting it right the first time*, Doubleday (N.Y)
10. Evans, J.St B.T. (1990) Bias in human reasoning: causes and consequences. In *Essays in Cognitive Psychology*, Lawrence Erllbaum Associates, Hove and London (paperback) p. 118
11. Segall, M.D. (2008), Why is it still drug discovery? *European Biopharmaceutical Review*, Spring issue.
12. Cuatrecasas, P. (2006) Drug discovery in jeopardy, *J. Clin. Invest.* 116, 2837-2842
13. Sutherland, S. (2007) *Irrationality*, Pinter and Martin (paperback)
14. Goldacre, B. (2008) *Bad Science*, Fourth Estate (paperback), additional chapter available free of charge through http://www.amazon.co.uk/Bad-Science-Ben-Goldacre/dp/000728487X/?tag=bs0b-21
15. Kahneman, D. and Lovallo, D. (1993) Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science* 39, No. 1, 17-31
16. Seligman, M.E.P. *et al* (2006) Positive psychotherapy. *American Psychologist*. 61(8), 774-788
17. Segall, M.D. *et al* (2006) Focus on Success: Using in silico optimisation to achieve an optimal balance of properties. *Expert Opin. Drug Metab.Toxicol.* 2, 325-337

18. Graber, M.L. *et al* (2005) Diagnostic Error in Internal Medicine. *Arch. Intern. Med*. 165, 1493-1499.
19. Reference 6 p.127
20. Bonabeau, E *et al* (2008) A more rational approach to new-product development. *Harvard Business Review* March, 96-102
21. Feuerstein, G.Z. and Chavez, J. (2009) Translational Medicine for Stroke Drug Discovery: The Pharmaceutical Industry Perspective. *Stroke* 40, S121-S125.
22. Gladstone, D.J. *et al*. (2002) Toward wisdom from failure: lessons learned from neuroprotective stroke trials and new therapeutic directions. *Stroke* 33, 2123-2136
23. Sena, E.S. *et al* (2010) *PLoS Biology* 8, e1000344
24. Feuerstein, G.Z. *et al* (2008) Missing steps in the STAIR case: a Translational Medicine perspective on the development of NXY-059 for treatment of acute ischemic stroke. *Journal of Cerebral Blood Flow & Metabolism* 28, 217–219
25. Grotta, J. (2002) Neuroprotection Is Unlikely to Be Effective in Humans Using Current Trial Designs. *Stroke* 33, 306-307
26. Elstein, A.S. and Schwarz, A. (2002) Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ* 324, 729-732
27. Bradley, C.P. (2005) Can we avoid bias? *BMJ* 330, 784
28. Esserman, L. *et al* (2002) Improving the Accuracy of Mammography: Volume and Outcome Relationships. *JNCI* 94(5), 369-375
29. Goodwin, P. and Wright, G. (2004) *Decision Analysis for Management Judgment* (3rd edn), Wiley, p.270
30. Gonzalez, C. and Brunstein, A. (2009) Training for Emergencies. *J Trauma* 67(2) S100-S105
31. The answer is (c), approximately 10%. Of 1000 compounds, 108 (990 x 0.1 + 10 x 0.9) would be reported as toxic by the test, of which only 9 really are toxic.
32. Gigerenzer, G. and Edwards, A. (2003) Simple tools for understanding risks: from innumeracy to insight. *BMJ* 327, 741-744
33. DeWitte, R.S. (2002) On experimental design in drug discovery *Curr. Drug Disc.* 2, 19-22
34. Hardman, D. (2009) *Judgment and Decision Making: psychological perspectives*, BPS Blackwell, Chapter 7.
35. Reference 29 pp. 449-451.
36. Gittins, J. (1996) Quantitative methods in the planning of pharmaceutical research. *Drug Info Journal* 30(2), 479-487
37. Segall, M.D. *et al.* (2009) Beyond Profiling: Using ADMET models to guide decisions. *Chem. Biodiv*. 6(11), 2144-2151
38. Segall, M.D. *et al.* (2009) Guiding the Decision-Making Process to Identify High Quality Compounds. *Drug Metabolism Reviews* 41(s3), 7-186 Abstract 244
39. Ades A.E., Welton N.J. *et al.* (2008) Multiparameter evidence synthesis in epidemiology and medical decision making. *J. Health Serv. Res Policy* 13, 12-22
40. Chadwick, A.T .and Edwards, R.A. (2009) Smart planner: radically accelerating R&D to combat a biothreat. *Drug Dev. Res.* 70(4), 335-348
41. Sim, I. *et al.* (2002) Evidence-based practice for mere mortals: the role of informatics and health services research. *J Gen. Intern. Med* 17, 302-308
42. Wason, P.C. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology,* 12, 129-140