Scientific objective: In a recent blog post Dr Andrew Thorburn \(^1\) asks why many published discoveries cannot be replicated. He summarizes examples from Dr Glenn Begley (Vice President for oncology at Amgen) that show how Amgen has been unable to reproduce many high-profile discoveries. Dr Thorburn raises the possibility that ‘sloppy science’ may partly explain our inability to measurably improve survival in many cancers. Dr Thorburn reminds us of the importance of careful laboratory science such as validated antibodies, correct reagents, and adequate controls, and calls on the research community to ‘raise our game’.

It is imperative that we eschew bad practice in all research. The bigger challenge, and the greater opportunity for success is to reconceptualize our scientific paradigm for discovering new therapies, targeted or otherwise. The recent movie *Moneyball* describes such a reconceptualization in professional baseball; the idea was actually developed by a statistician, Bill James, and implemented by Oakland Athletics manager Billy Beane in the 2002 season. The result has changed the paradigm for building teams throughout professional baseball. As explained by the fictional Peter Brand:

> “Your goal shouldn’t be to buy players. Your goal should be to buy wins and in order to buy wins, you need to buy runs.”

The paradigm shift was not well received by the baseball establishment; as stated by a senior scout:

> “Baseball isn’t just numbers, it’s not science. If it was then anybody could do what we’re doing, but they can’t because they don’t know what we know. They don’t have our experience and they don’t have our intuition.”

But ultimately the A’s changed the way teams were developed, as stated by the owner of the Boston Red Sox as he tried to recruit Billy Beane away from Oakland:

> “The Oakland A’s won the exact same number of games that the Yankee’s won, but the Yankee’s spent one point four million per win and you paid two hundred and sixty thousand.”

So can we, like Oakland, find a system for discovering new therapies for one-fifth the cost in time and patients? When developing therapeutics, a “win” occurs when a new therapy that provides important benefits is adopted into routine practice. In contrast, adopting a non-beneficial or harmful therapy is a definite loss.

**Formulating the scientific objective:** From both a public health and scientific perspective we want a high probability of declaring that a good treatment is good, and importantly, a low probability of finding that a bad treatment is good.

\(^1\) [http://www.coloradocancerblogs.org/target-cancer/thorburn-better-treatment-for-cancer-start-by-ending](http://www.coloradocancerblogs.org/target-cancer/thorburn-better-treatment-for-cancer-start-by-ending)
This objective is formalized if we view clinical trials as diagnostic tests to identify good and bad treatments. A positive test is represented by a statistically significant trial which indicates that a new treatment has significant benefits. The accuracy of the diagnostic test is quantified by the familiar operating characteristics, which also apply to clinical trials:

- **Sensitivity:** The statistical power (which I call $\beta$) is the chance that the experiment correctly identifies a beneficial treatment (i.e., the chance of a statistically significant result if in fact the alternative hypothesis is true).

- **Specificity:** The familiar $\alpha$ level corresponds to specificity: $1 - \alpha$ is the chance that our experiment will correctly identify ineffective treatments; that is, $1 - \alpha$ is the probability that we do not find statistical significance when the null hypothesis is true.

- **Prevalence ($\pi$):** Instead of the prevalence of disease in the population, this is the prevalence of truly beneficial therapies among all therapies that might be considered (i.e., the pre-test probability). In fact, this prevalence is quite low. The NCI Developmental Therapeutics Program has evaluated over 400,000 candidate compounds since 1955 (with over 80,000 screened since 1990). The NCI is sponsoring approximately 1500 clinical trials that enroll over 25,000 patients each year. It has been reported that only 10% of drugs entering phase I trials are eventually found to be positive in subsequent phase III trials [1]; a systematic review of NCI-sponsored phase III trials from 1955-2006 [2] found that of 743 randomized comparisons, only 176 (24%) found statistically significant benefits from new therapeutics. As systematic review of phase II trials [3] found that 39 of 266 (15%) led to phase III trials. Thus, the prevalence of truly beneficial therapies among all treatments entering phase II trials is probably less than 10%.

**Public health objective:** *high Positive predictive value (PPV).* Among all positive “diagnoses” (statistically significant trials) we want a high prevalence of truly beneficial treatments and a low prevalence of non-beneficial treatments. As with clinical diagnostics, the positive predictive value can be calculated from the characteristics of the experiment:

$$PPV = \frac{\beta \pi}{\beta \pi + \alpha (1 - \pi)}$$  

Returning to Dr Thorburn’s observation, scientific inquiry will never be 100% sensitive and specific, there are always false positives. The key to success is to recognize this important principle, and as we now discuss, reduce the false positive risk through replication, good practice in clinical trials, and good trial design, with the ultimate goal of higher PPV.

**Scientific objective illustrated:** This abstract construction (see Ioannidis [4] for additional discussion) can be illustrated with a specific example. Suppose that a cooperative group has the resources to evaluate 2000 new therapies. Standard early phase (think phase IIb) trials are used, with statistical significance defined as a $p$ value less than 0.05. Suppose that power for the clinical endpoint is 40% (which is probably high for clinical endpoints such as mortality in a phase IIb trial). Now suppose that only 10% of these new therapies are in fact efficacious (i.e., 200 beneficial
and 1800 ineffective therapies). On average after completing 2000 trials we would see:

\[
\begin{align*}
200 \times 0.4 & = 80 \text{ true positive results} \\
1800 \times 0.05 & = 90 \text{ false positive results}
\end{align*}
\]

The positive predictive value is only \(80/170 = 47\%\).

Now, when the cooperative group (or pharma) tries to replicate the significant results in the 170 therapies in larger phase III trials (using 80% power); we get the following result:

\[
\begin{align*}
80 \times 0.8 & = 64 \text{ true positive results} \\
90 \times 0.05 & = 4 \text{ false positive results}
\end{align*}
\]

So that ultimately, \(64/68 = 94\%\) of what is recommended for use in the population is in fact efficacious. Notice how the phase II screening trial increased the prevalence of truly beneficial therapies for subsequent phase III trials from 10% to 47%. Phase III trials are conducted in an enriched set of treatments so that the ultimate PPV is over 90%. As shown in Appendix Example 1, the use of phase II screening trials before phase III trials gives high PPV and more beneficial treatments than proceeding with phase III without phase II trials.

**Common pitfalls that reduce scientific efficiency:** I now describe three common practices that hurt our win-percentage and decrease the efficiency with which new therapeutics are discovered. Changing these practices would change the paradigm and facilitate the discovery of truly effective therapies more rapidly and with fewer patients.

1. **Negative results are scientifically important:** A well-designed experiment needs to be informative regardless of its outcome. Lack of statistical significance is often viewed as failure - it shouldn’t be if the experiment is properly designed; it is equally important to discover what is not true. Unfortunately academic biomedical research values positive results more than negative results. Thus, statistically significant beneficial results often determine whether your paper is accepted in a high-impact journal, and whether your next grant gets funded. Therefore the Darwinian forces in academic biomedicine select for very common detrimental scientific practices all of which increase the false positive rate (\(\alpha\) level) and reduce specificity. Examples include:

   - Analyzing multiple outcomes: if the primary analysis is not significant, find a different way to analyze the data.

   - Subgroup analyses: Evaluate whether the results are significant in subgroups (e.g., those defined by genotype or biomarker target). This is becoming a bigger problem as we seek targeted therapies and personalized medicine.

   - Surrogate endpoints: Endpoints such as biomarker level, tumor response, or time to progression are common surrogate endpoints in cancer trials. There are many examples of false positive treatments that are beneficial on a surrogate endpoint but lack benefit or are harmful on the primary clinical endpoint.

   - Publishing interim results: frequent interim analyses increase the risk of a false positive. We should not present interim results in abstracts or use that information in grant applications.
• Inadequate controls: Uncontrolled clinical trials have a well-earned reputation for producing overly-optimistic evidence for treatment benefit. A concurrent randomized control group is necessary to obtain unbiased estimates of treatment effect. The experience of many statisticians is captured in a common quote: “There are two kinds of investigators: those with lots of enthusiasm but no controls and those with lots of controls but no enthusiasm.”

Although biomedical researchers understand the dangers of multiple-comparisons and the attendant increase in the risk of false positive conclusions, very few seem to accept the approach illustrated above in which no means no and treatments that are negative in a phase II trial are dropped from further consideration. Instead there is a worrisome view that “there are no bad anticancer agents” [1]; that is, exploratory analyses and minor treatment modifications are justified because every biologically-promising candidate treatment might actually be a cure – we just have to find the right disease. Appendix Example 2 explores this never give up paradigm and the attendant effects of increasing the $\alpha$ level in equation (2); for example if $\alpha = 0.34$ instead of 0.05, then ultimately 64 beneficial and 31 non-beneficial treatments are declared effective (PPV = 67%) in phase III trials. In addition to reducing PPV from 94% to 67%, this approach requires 692 instead of 170 phase III trials.

A related approach is illustrated in Appendix Example 3 (try try again) in which multiple analyses are conducted at the end of the phase II trials, but significant findings are confirmed in a subsequent phase II trial before a phase III trial is considered. This example produces an ultimate PPV which is similar to that of example 1, but the number of patients required is larger and fewer beneficial treatments are discovered.

2. The importance of adequate power, even in early phase trials: The examples above assume 40% power in a phase II trial. If instead power is reduce to 20% in equation (2) then phase II PPV is reduced from 47% to 31% with the number of beneficial treatments discovered in phase III reduced by half from 64 to 32 and ultimate PPV reduced from 94% to 88%. Commonly, phase II trials do not even consider the power for the ultimate clinical endpoint (e.g., survival). For example, it is common to use a surrogate endpoint of tumor response or risk of progression when ultimately a mortality endpoint is warranted. In this situation, the sample size is often so small that there is almost no information (very low power) for the clinical endpoint. As discussed above, surrogate endpoints can also raise the false-positive risk; in fact if surrogates simultaneously increase $\alpha$ and reduce $\beta$ so that the power is less than the $\alpha$ level ($\alpha \geq \beta$) then by equation (1) $PPV \leq \pi$; thus, phase II trials with non-clinical endpoints might actually increase the percentage of non-beneficial treatments studied in phase III trials to more than the pre-evaluation prevalence ($\pi$). We should consider designing phase II trials with low power for the clinical outcome as opposed to high power for a surrogate outcome. Note that phase II trials may provide essential information on adverse effects and on mechanistic information about treatment effects on a target. We acknowledge that there are always multiple scientific interests in clinical trials, however they must be scientifically prioritized in order to control the operating characteristics that determine PPV.

3. “Novel” and “Innovative:” the double-edged sword of science: The disappointing therapeutic success rate in many cancers has increased the call for novel and innovative therapies. “Novel” and “innovative” cut both ways: while we hope for discovery of a magic bullet, the prevalence of such breakthrough discoveries is clearly quite rare. If the prevalence is
reduced to 1% ($\pi = 0.01$) in the illustration of equation (2), then phase II studies will discover 8 breakthrough therapies and 99 non-beneficial therapies (PPV = 7.5%). Subsequent phase III trials yield 6 beneficial and 5 non-beneficial therapies for an ultimate PPV of about 55%. Once again invoking Darwin, the selective factors in academic biomedicine favor funding and publication of novel and innovative treatments. The benefits to patients may in fact be larger if we instead focus on improving existing treatments through studies of dosing schedules and ancillary care. Unfortunately, incremental advances are less appealing to funding agencies than novel ideas.

Note that the same concern applies to novel and innovative approaches to the statistical design of clinical trials. Although there is renewed interest in adaptive designs, some of these designs do not use decision criteria that control the operating characteristics which govern PPV in equation (1); it is particularly concerning if these designs increase the $\alpha$ level. **Appendix**

**Example 4** illustrates how adding interim analyses to a trial design can reduce the patient requirements by 34.6% when compared with an equivalent design without interim analyses. As long as the interim analyses are planned to maintain the same $\alpha$ level and power, then these efficiency gains will not affect the PPV.

**Concluding remarks:** We can indeed improve the efficiency of biomedical research and the PPV for the discovery of treatments that have real benefit to human health. The basic principles of practice are represented by the key components discussed above:

(a) **Control specificity:** Study meaningful and informative questions, and ensure that the design provides information regardless of the outcome: negative results should be as informative as positive. Control the false positive risk ($\alpha$ level) by avoiding multiple analyses of many endpoints and subgroups. Any interesting discovery in early-phase trials must be studied in subsequent early-phase trials (avoid direct progression to phase III).

(b) **Increase power:** Not only with sample size, but also with good experimental practice and good recruitment and retention. Reduce variability with central adjudication of endpoints and eligibility. Eschew sloppy science because it commonly increases variability or introduces bias and leads to unreproducible results.

(c) **Improve prevalence:** Previous experience [1][2][3] indicates that truly beneficial treatments are rare. In this situation, rather than focusing on novel treatments, greater public health benefits might be achieved with incremental advances in dosing schedules or ancillary care. Better understanding of mechanism and treatment targets (especially in pre-clinical studies) might improve the information in early-phase trials so that ineffective treatments can be eliminated from costly further study in either phase II or III trials.

Finally, improvements in the efficiency with which we discover new treatments is both necessary and possible as long as we keep our eye on the prize and structure our science toward the objective of a higher win percentage.
Appendix: Illustrative Examples

Example 1 *(No means no, and the efficiency of screening trials)*: The example of equation (2) illustrates the general principle that phase II screening trials improve the efficiency of discovering beneficial therapies. The example used patients for 2000 phase II trials and 170 phase III trials. To raise power from 40% to 80% takes 2.7 times more patients; thus, each phase III trial takes about 2.7 times the sample size of the phase II trials. Therefore the total sample size for the cooperative group development program is equivalent to 2459 phase II trials (2459 = 2000 + 170 × 2.7). The same number of patients could have instead been used to study 910 (= 2459/2.7) treatments in phase III trials (without first testing in phase II trials); these two scenarios illustrate the general principle that phase II screening trials improve efficiency:

*Phase III only:* Suppose that instead of phase II screening trials the cooperative group uses all of their resources in 910 phase III trials with power \( \beta = 0.8 \) and \( \alpha = 0.05 \). If the prevalence is 10% then they will test 91 beneficial treatments and 819 non-beneficial treatments. At the end of the phase III program they will discover 73 beneficial treatments (= 91 × 0.80) and 41 non-beneficial treatments (= 819 × 0.05) with an ultimate PPV of 73/114 = 64%.

*Phase II screening followed by phase III trials:* As shown in equation (2) phase II screening trials improve the prevalence of beneficial treatments from 10% to 47% so that phase III trials are conducted in this enriched population. Ultimately using screening trials leads to discovery of 64 beneficial treatments and 4 non-beneficial treatments (PPV = 94%).

Example 2 *(Never give up)*: There are many reasons why clinical trials of new therapies might not be stopped after a single phase II trial. The arguments are probably best articulated by Von Hoff [1], who worries that programmed drug death (pharmacoptosis) leads to premature (and inappropriate) termination of clinical trials on many biologically promising therapeutics. From this perspective, phase II trials need to be fully analyzed to determine what works; specifically, are there subgroups (defined by treatment targets, genotype, or other characteristics) in which the treatment worked, or are there endpoints that demonstrate efficacy? The challenge for scientific efficiency is that the multiple analyses required to explore all possible results lead to larger values for \( \alpha \) and an excess of false positive results. To illustrate, suppose that the decision to continue into phase III trials is based on whether any of 8 independent statistical analyses is significant; such an approach will increase the \( \alpha \) level to 0.34. Using \( \alpha = 0.34 \) in the illustration of equation (2) gives 612 false positive and 80 true positive phase II trials (phase II PPV = 0.12), which in turn has two important consequences for scientific efficiency:

- If phase III trials are conducted on all 692 positive phase II treatments, then ultimately 64 beneficial and 31 non-beneficial treatments are discovered. As a result PPV is reduced from 94% to 67% (64/95).
- Increasing \( \alpha \) from 0.05 to 0.34 increases the number of phase III trials from 170 to 692. This increase in the number of phase III trials is equivalent to conducting 3868 phase II trials (= 2000 + 692 × 2.7); i.e., a 57% increase (=3868/2459) in time and number of patients over example 1. In order to avoid increasing the number of patients, the cooperative group could reduce the number of phase II studies from 2000 to 1270, which would then lead to 439 phase III trials. Now the phase III trials would find 41 beneficial and 19 non-beneficial
treatments. Although the PPV remains at about 67% the number of beneficial treatments that are discovered is reduced from 64 to 41.

In fact multiple comparisons can increase the \( \alpha \) level beyond 0.34. Using \( \alpha = 0.5 \) in equation (2) identifies 80 beneficial and 900 non-beneficial treatments in phase II trials, which means that the phase II trials have actually increased the prevalence of ineffective treatments in the phase III trials (PPV = 8.2% which is smaller than \( \pi = 10\% \)). As a consequence, phase III trials discover 64 beneficial and 45 ineffective treatments (ultimate PPV = 59%) and require the equivalent of 4646 phase II trials.

**Example 3 (Try try again):** One approach to addressing the scientific deficiencies illustrated in the never give up paradigm is to allow multiple analyses in the initial phase II trial, but to require that all significant phase II treatments be replicated in a second phase II before phase III trials are considered. To illustrate suppose that 2000 treatments are screened, that 1800 of these are ineffective, that 100 are beneficial on the primary comparison, and that 100 are beneficial in a subgroup or by endpoints that might be discovered in secondary analyses. The result of the initial 2000 phase II trials (with \( \alpha = 0.34 \) and \( \beta = 0.4 \)) is that 200 \times 0.4 = 80 truly beneficial treatments will be discovered in a combination of primary and secondary analyses, and 1800 \times 0.34 = 612 ineffective treatments will be significant in primary or secondary analyses. These 692 treatments will then be studied in a second round of phase II trials using the endpoints and populations that were significant in the initial phase II. The second phase II trial will not allow multiple comparisons by controlling \( \alpha = 0.05 \) and \( \beta = 0.4 \). The second phase II trials yield 80 \times 0.4 = 32 truly beneficial treatments, and 612 \times 0.05 = 31 false positive treatments. The subsequent phase III trials (with \( \alpha = 0.05 \) and \( \beta = 0.8 \)) yield 32 \times 0.8 = 26 beneficial treatments and 31 \times 0.05 = 2 ineffective treatment (for ultimate PPV = 0.93). Notice that this paradigm requires a total sample size for an equivalent of 2862 phase II trials (2862 = 2000 + 692 + 63 \times 2.7), which represents an 16% increase (2862/2459) over the paradigm of example 1; not only does it require more patients, but the number of beneficial treatments discovered is reduced from 64 to 26. From a practical viewpoint: If multiple rounds of phase II trials are required, then the total number of patients and time increases accordingly. Also, if treatments are targeted toward increasingly narrow patient populations or disease indications, then recruitment becomes more difficult and the public health impact also decreases.

**Example 4 (Efficiency in trial design):** Recent interest in innovative trial designs is motivated by the need for more efficient drug discovery. Adaptive designs allow mid-trial changes in endpoints, sample size, or even patient populations. As long as these designs are planned and implemented to control the trial’s \( \alpha \) level and power, then these designs are very similar to classic group sequential designs that allow early termination of clinical trials. If the \( \alpha \) level and power are controlled, then an adaptive design will maintain PPV as illustrated in example 1. On the other hand, if these operating characteristics are altered by the design then the PPV will also be altered as illustrated in examples 2 and 3. The primary benefit of sequential clinical trials is that early termination allows decisions with a smaller sample size. For example, the one-sided symmetric design of Emerson and Fleming [5] using 5 interim analyses and O’Brien-Fleming stopping rules allows early termination for either lack of effect or for evidence of benefit. The expected sample size for such a design is only 65.4% of an equivalent fixed-sample design. Thus, the 2000 phase II trials and 170 phase III trials would need only 65.4% of the patients because many of the trials would stop early for either lack of effect or for efficacy. Thus, appropriate trial monitoring improves efficiency through a 34.6% reduction in patient number with attendant reduction in time and financial costs.
References


