A reprint from
American Scientist
the magazine of Sigma Xi, The Scientific Research Society

When Averages Hide Individual Differences in Clinical Trials

Analyzing the results of clinical trials to expose individual patients' risks might help doctors make better treatment decisions

David Kent and Rodney Hayward

In 1793, when yellow fever reached Philadelphia, killing hundreds of people, nobody knew its cause. But there was no shortage of theories. Based on these, a few desperate physicians devised increasingly radical treatments. One of the most famous, concocted by the renowned Dr. Benjamin Rush, was the "Ten-and-Fifteen" purge, a combination of 10 grains of calomel, a mercury-based compound, and 15 grains of jalap (the poisonous root of a Mexican plant related to the morning glory). After administering this toxic brew, doctors then bled patients so profusely they often passed out. Miraculously, a few hearty patients survived both the sickness and its cure, reinforcing the doctors' belief in the value of their treatment.

The confusion of Rush and his contemporaries is understandable; given the natural variation in patient outcomes, it can be surprisingly difficult to tell whether a treatment is helping or harming patients. Things are no different today: Therapies that are useless or worse can still inspire great enthusiasm among physicians. But unlike Rush, modern physicians are at least somewhat protected from the human tendency to draw unwarranted conclusions by a statistical instrument called the clinical trial.

The most powerful form of trial, the randomized controlled clinical trial, was devised as a means of determining a treatment's effect when many other factors, including unknown ones, might affect patient outcomes. Over the past 50 years, large-scale human trials have paid rich dividends in lives saved and improved quality of life. In 1972, a Scottish epidemiologist named Archie Cochrane published a book urging physicians to follow the evidence of clinical trials in their practices. By the 1990s a group of doctors led by the Canadian physician David Sackett had coined the term "evidence-based medicine," and a movement was born. These physicians advocated that everyday treatment decisions be guided by the results of systematic reviews of clinical trials.

What could be more reasonable? And yet precisely because evidence-based medicine gives impersonal statistical data greater weight than clinical experience, it has met strong and at times emotional resistance from practicing physicians. Some see this resistance as a self-interested reaction, but we believe that it arises in part from a fundamental mismatch between the evidence provided by clinical trials and the needs of practicing doctors treating individual patients. Because many factors other than the treatment affect a patient's outcome, determining the best treatment for a particular patient is fundamentally different from determining which treatment is best on average.

We believe that changes in the way clinical trials are analyzed could offer at least a partial solution to this dilemma and yield the more detailed information doctors need to make better treatment decisions.

The Modern Clinical Trial

The clinical trial is a surprisingly recent invention. The first modern trial, conducted in 1947–48, showed that the newly discovered antibiotic streptomycin was more effective than the conventional treatment for tuberculosis. It was to be 15 years, however, before drugs routinely underwent clinical trials prior to being sold in the United States. In the late 1950s, severe birth defects were reported after the tranquilizer thalidomide was given to pregnant women. This tragedy spurred Congress to pass the Kefauver-Harris Drug Amendments of 1962, which finally forced manufacturers to prove that a new drug was both safe and effective.

The randomized controlled clinical trial became the standard means of providing this proof. The patients in this type of trial are assigned to one of two groups—the experimental

David Kent and Rodney Hayward are both researchers and practicing physicians specializing in general internal medicine. Kent is an assistant professor of medicine at Tufts-New England Medical Center, Tufts University and also associate director of the clinical research program and assistant professor of clinical research at the Sackler Graduate School of Biomedical Sciences at Tufts. Hayward is the director of the Veterans Affairs Ann Arbor Health Services Research & Development Center of Excellence and is a professor of public health and internal medicine at the University of Michigan. He was the 2005 recipient of the Under Secretary for Health Award for career accomplishments in health-services research. Address for Kent: Institute for Clinical Research and Health Policy Studies, 750 Washington Street, Tufts-NEMC #63, Boston, MA 02111. Internet: dkent1@tufts-nemc.org

Figure 1. In the year 2000 artist Chris Dorley-Brown took 2,000 digital photos of people living in the small town of Haverhill in Suffolk, England, and then used software to merge these photographs, step by step, until all 2,000 had been blended into one image. This illustration shows 12 double portraits (two photographs morphed together) and a blend made up of all of the double portraits. (Dorley-Brown has left out the original portraits in order to protect the privacy of the participants.) In the modern clinical trial, the responses to treatment of thousands of individuals are typically summarized in a single number in the same way the center photograph represents all the other individuals. As the data are averaged, important individual differences are lost. The authors propose ways to better examine clinical-trials results to guide medical practice.

group or the control group—and assessment of the outcome measure is typically blinded or masked (the assessing physician does not know whether patients received treatment). Randomization is the feature that gives trials the power to find the treatment's effect in the clutter of different patient risk profiles. If patients are randomly assigned to the experimental and control groups, risk factors should be equally distributed between the groups. Thus any difference in the aggregated outcomes of the two groups can be attributed to the effects of treatment.

The treatment-effect, as it is called, is typically a single number that summarizes the overall result of the trial. The treatment-effect can be expressed as the absolute risk reduction (the difference between the outcome rate in the experimental group and the outcome rate in the control group) or the relative risk reduction (the decrease in bad outcomes in the experimental group relative to the outcome rate in the control group). The absolute risk reduction is always a much smaller number than the relative risk reduction. For example, if a trial shows that a statin drug decreases the risk of heart attacks from 6 percent (the outcome rate in the control group) to 4 percent (the outcome rate in the experimental group), the absolute risk reduction

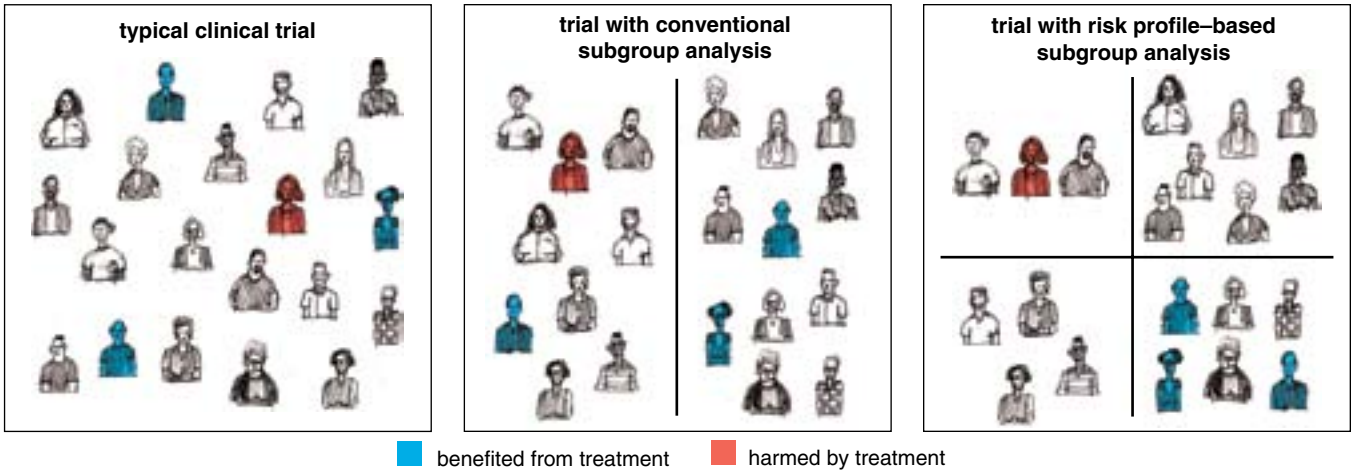


Figure 2. The patients in a large clinical trial inevitably have different health histories and risk factors. Many medical investigators have assumed that this kind of heterogeneity is an advantage because it makes the trial population more like the real population. But summarizing treatment results for a population in a single number may gloss over subgroups of patients whose response is quite different from the statistical average. Conventional "one variable at a time" subgroup analysis is unlikely to find subgroups with large differences in response to therapy. Newer types of analysis, using risk profiles, may be more powerful in grouping patients according to their likelihood of benefiting from treatment.

is 2 percent and the relative risk reduction is 33 percent (the absolute risk reduction divided by the outcome rate in the control group).

Doctors are more likely to adopt a treatment when the treatment-effect is expressed using a larger, more impressive number, even though the information underlying calculations of absolute and relative risk is identical. Thus, trial sponsors (frequently pharmaceutical companies) typically emphasize the larger relative risk reduction. But whichever way treatment-effect is expressed, reporting a single number gives the misleading impression that the treatment-effect is a property of the drug rather than of the interaction between the drug and the complex risk-benefit profile of a particular group of patients. Consider what happens when sicker patients are enrolled and the rate of the problematic outcome in the trial goes up. If the relative risk reduction stays the same, the absolute benefit must get proportionally larger. This reflects our intuition that sicker patients have potentially more to gain from therapy.

But when treatments have even a small risk of serious harm, the differences in treatment-effect may not just be a matter of degree. Indeed, some patients may benefit substantially from a treatment even when the overall results from a trial are negative. Or a treatment with benefit on average may be extremely unlikely to help most patients, while being more likely to harm than help some others. But unless the trial investigators analyze their data looking for these subgroups, the physician cannot know whether they exist.

Hidden Risks

Harm to a few was the problem lurking in the statistics of the landmark GUSTO study, which compared two thrombolytic (clot-busting) drugs for heart-attack victims. In the 1970s several drugs were found that could dissolve a clot and restore blood flow to heart muscle before it was irretrievably damaged. One of these was streptokinase. But in 1978 a Belgian scientist discovered that the cells lining blood vessels made an enzyme, tissue-type plasminogen activator, or t-PA, that also dissolved clots. In the early 1990s the biotechnology company Genentech, which had succeeded in genetically engineering this enzyme, and the National Institutes of Health sponsored a huge clinical trial of streptokinase and t-PA.

The trial showed that t-PA was considerably more effective than streptokinase, reducing the relative risk of death by about 15 percent. The newer drug was also much more expensive than streptokinase, but analysis showed that its benefits justified the additional expense. Following the GUSTO study, use of streptokinase declined dramatically, and it is now very rarely used for heart attacks in the U.S.

Of course, t-PA does not reduce every patient's risk by the same amount. Consider, for example, two patients who both qualify for thrombolytics. Estragon is 72 and diabetic. When he arrives at the emergency room by ambulance, he is experiencing severe chest pain and has a rapid pulse and low blood pressure. An electrocardiogram indicates a heart attack affecting a large and vital area of the heart muscle. Vladimir, 52, has stable vital signs and no chronic illnesses. He has come to the emergency room complaining of chest pressure. His electrocardiogram indicates that he has had a heart attack affecting only a small area of the heart muscle.

Given his condition, Estragon's mortality risk without thrombolytics would be about 25 percent, whereas Vladimir's would be close to 2 percent. Estragon is at such high risk of dying that the potential benefits of t-PA clearly outweigh any risks or costs associated with this

agent. But it is not clear that t-PA would benefit Vladimir, who is highly likely to survive no matter which thrombolytic agent he receives. In fact, if Vladimir has high blood pressure or a history of stroke, both of which would increase his risk of intracranial bleeding, giving him the more potent t-PA might actually *increase* his risk of dying (albeit only slightly).

In the GUSTO trial lower-risk patients like Vladimir were much more common than higher-risk patients like Estragon. When we re-analyzed the GUSTO results using mathematical models that estimated the risk of death based on patient characteristics, we discovered that t-PA primarily benefited a subgroup of high-risk patients. The highest-risk quartile of patients accounted for most of the outcomes that gave t-PA the edge over streptokinase. Paradoxically, even though the overall results of the trial suggest that t-PA is better and clearly worth the extra risks and costs, the benefits for the typical patient in the trial are less, and the trade-offs less clear.

Hidden Benefits

Summarizing trial results may exaggerate the benefit of treatment for some patients, but the reverse is also possible: A negative overall result can hide significant benefit to some patients. Consider, for example, the ATLANTIS B trial, undertaken in the late 1990s. This trial tested the efficacy of t-PA in treating strokes instead of heart attacks. Strokes are trickier than heart attacks because the thrombolytics must be given much sooner (within 3 rather than 12 hours) and the risk of thrombolytic-related intracranial hemorrhage is much greater (probably 6 or 7 percent instead of 1 percent).

Earlier clinical trials had shown that t-PA did not yield any overall benefit if it was administered more than three hours after the patient first had symptoms of a stroke. This short window of opportunity meant t-PA was given to fewer than 5 percent of stroke patients. To re-test the treatment window, ATLANTIS B enrolled patients arriving for treatment between three and five hours after the onset of symptoms of a stroke. The trial demonstrated no overall benefit for t-PA (treated patients were no more likely than those who received a placebo to recover normal or near-normal function). Moreover, as mentioned above, treatment with t-PA substantially increased their risk of intracranial hemorrhage.

Physicians looking only at the average result of this trial would be understandably discouraged by the lack of benefit and the increased risk of harm. But the trial showed that t-PA and placebo were essentially equivalent. The fact that some patients given t-PA were harmed by it implies others must have benefited from it. We hypothesized that if patients at lower risk of intracranial hemorrhage could be identified and t-PA given only to them, the treatment-effect might be different. When we used a risk model derived from independent data to divide the ATLANTIS B patients into thirds, we found that the third of the patient population at the lowest risk of thrombolytic-related hemorrhage actually did better with t-PA—even though they were treated outside the approved time window.

The paradoxical results of the GUSTO and ATLANTIS B trials arise from underlying variation in the baseline risks of these populations. For GUSTO, variation in the degree of benefit was due mostly to large variation in the risk of the outcome (death). For ATLANTIS B, it was attributable to variation in the risk of treatment-related harm.

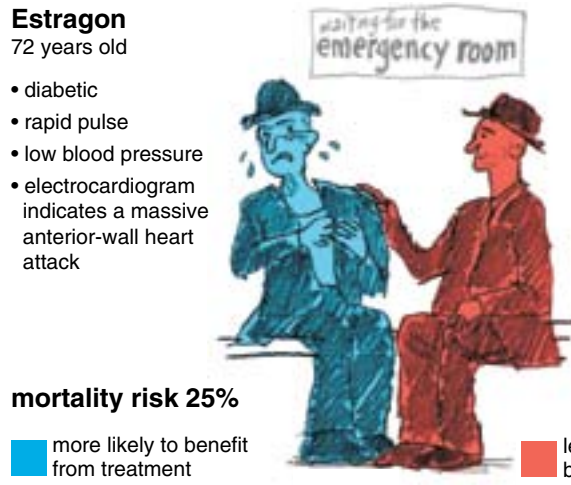


Figure 3. The common practice of treating everyone just to make sure patients who will benefit get treated is dangerous if a treatment carries a risk of harm for some patients. Two hypothetical heart-attack patients illustrate this situation. Estragon is very ill and would probably benefit from t-PA, a clot-busting drug that performed better than streptokinase in a randomized clinical trial. Vladimir, a patient at low risk of dying from his heart attack, might actually be harmed by the more potent agent, particularly if he has risks for bleeding, such as high blood pressure or a prior stroke.

John Ioannidis and Joseph Lau, our colleagues at the University of Ioannina in Greece and Tufts-New England Medical Center respectively, have advocated measuring the degree of variation in outcome risk in a trial by comparing the outcome rate in the quarter of patients with the lowest risk score to the outcome rate in the quarter with the highest risk score. In the GUSTO trial, the mortality rate in the highest-risk quartile is nearly 10 times higher than that in the lowest risk quartile.

This degree of variation may seem high, but it is not extreme by any means. Looking at trials of treatments for HIV infection, Ioannidis and Lau found examples where the outcome rate in the high-risk group was more than 50 times higher than that of the low-risk group. And when we looked at trials testing blood-pressure medicine for chronic kidney disease, we found similar ratios: Outcome rates were less than 1 percent in the low-risk patients and more than 30 percent in high-risk patients.

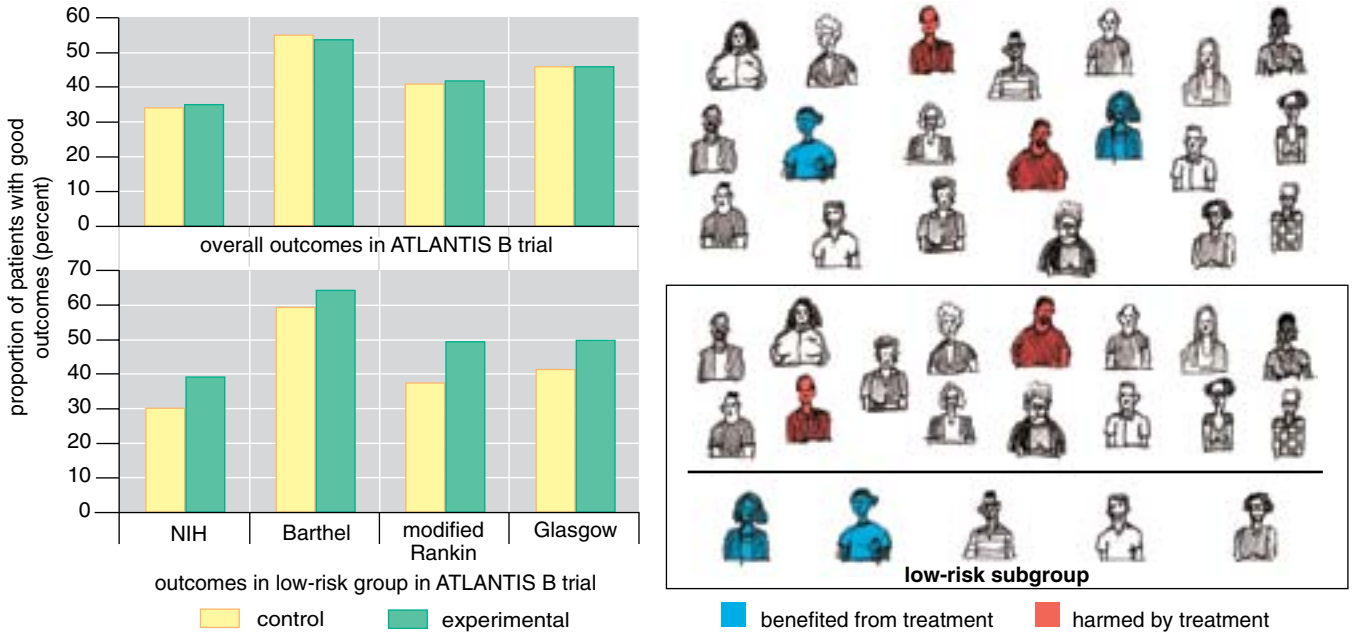


Figure 4. The ATLANTIS B trial looked at the outcome of stroke patients treated three to five hours after the onset of symptoms, slightly longer than the recommended cut-off. The collective results *(top graph)* indicated patients in the experimental group did no better than those in the control group, who were given a placebo. But when the authors looked at the outcomes among the one-third of patients at least risk of hemorrhage, they found that they were helped by t-PA *(bottom graph)*. In the initial analysis the benefit reaped by some was masked by the harm others suffered. Risk models that identify those patients least likely to suffer a treatment-related hemorrhage based on pretreatment characteristics *(right)* could potentially allow the treatment window for t-PA to be extended in appropriate situations. (Graph adapted from Kent, Ruthazer and Selker 2003.)

It should be obvious that when there is this degree of variation, one should not expect similar risk-benefit trade-offs in high and low risk groups. The high degree of variation arises because trials frequently enroll many patients with a negligible risk for the outcome even in the absence of treatment. For such patients, therapies associated with even a modest risk of treatment-related adverse effects will be causing net harm.

Not only is there considerable variation in risk, but it also appears that the baseline risk is not always distributed normally, in bell-curve fashion. If risk were distributed normally, then the overall trial result would at least reflect the outcome risk and treatment-effect in the typical patient. But in fact the GUSTO distribution, where many patients are at low risk and a few patients at high risk, might be more typical.

There are several reasons why risks might be highly skewed. One of them is something called the *floor effect*. Since there is no such thing as a negative risk, people with high outcome risks cannot be balanced out by people with risks less than zero. Another is that risk factors are not randomly distributed but instead clump: The presence of one risk factor often increases the likelihood of having others. When risks are skewed, the typical outcome risk may be different from the average outcome risk, and therefore the overall treatment-effect might not reflect the benefit to even the typical patient in the trial *(Figure 5)*.

One Variable or Many?

Many clinical trials include some attempt to explore differences in treatment-effect among the enrolled patients. But these analyses almost always focus on one attribute or risk factor at a time. For example, they might compare outcomes in men and women or in patients with and without hypertension. But one-variable-at-a-time subgroup analyses are not likely to yield meaningful information. For one thing, so many different variables can potentially influence the response to therapy and the likelihood of an outcome that if separate subgroup comparisons are made, chance alone will ensure that some subgroups show differences in treatment-effect. The hazards of "false positive" subgroup analysis from multiple comparisons were amusingly demonstrated using data from the ISIS-2 trial, which looked at the effect of aspirin in patients with heart attacks. Post-hoc subgroup analysis showed that aspirin did not lower mortality in heart-attack patients born under the signs of Libra and Gemini, but did in those born under other signs.

An equally important and less-well-appreciated reason that one-variable-at-a-time analysis is not effective is that a patient's outcome can be affected by many factors *simultaneously*. Since risks affect outcomes cumulatively, outcome differences between groups that differ by just a single risk factor tend to be relatively small. On the other hand, large outcome differences are found in analyses that compare subjects with many risk factors to subjects with none or few. Even if the experimenters pick a relatively strong risk factor, a single factor is unlikely to reliably discriminate between those who are at greatly different risks for the outcome and

who therefore have widely diverging risk-benefit trade-offs from treatment.

We believe that to be truly useful clinical trials must routinely include analyses that combine risk factors into risk scores or indices. Risk models for a wide variety of diseases can be found in the literature, although they have not yet been exploited to analyze clinical-trial results.

To demonstrate that our GUSTO and ATLANTIS B examples aren't anomalous and that multifactor analyses are by nature more powerful than single-variable analyses, we ran computer simulations of hypothetical clinical trials. We then grouped patients according to the presence or absence of one risk factor or according to a risk score based on their count of risk factors. When we looked at the outcomes of these simulated patients, single-factor subgroup analyses proved statistically weak; it was unlikely such analyses would reveal real and often large differences in the treatment-effect. Subgroup analyses using multiple factors, on the other hand, were extremely powerful; under typical circumstances, these analyses would reveal important differences in the treatment-effect in different risk groups. (See "Finding Answers for Vladimir and Estragon," *next page*.)

The risk scores in our simulation examined only one *dimension* of risk: the risk of having the outcome of interest. Important variation in this baseline risk is so common that analysis across this dimension should be routine. But the importance of other dimensions of risk should be explored in certain cases.

For treatments with a particularly high rate of serious adverse events, such as thrombolytics for stroke, scores for the risk of treatment-related harm may be helpful in discriminating patients likely or unlikely to benefit (as in our ATLANTIS B analysis). In some cases, there might be reason to examine characteristics that affect the relative responsiveness to therapy, such as time-to-treatment for thrombolytics or other emergency therapies—although, in this dimension, it might be hard to combine such characteristics into a score. Lastly, particularly for chronic diseases being treated over time in older, sicker populations, competing risks (the risk of succumbing to an illness not related to the treatment) could also give rise to differences in treatment-effect.

In any case, what is most important is that the myriad individual risk factors can be summarized into just a few risk dimensions, which are much more powerful than the individual variables in sorting patients into those likely and unlikely to benefit.

A Landmark Study

In 1999 Peter Rothwell of the Radcliffe Infirmary in Oxford, England, and Charles Warlow of Western General Hospital in Edinburgh, Scotland, published a landmark reanalysis that vividly shows how risk-benefit stratification can improve our understanding of a clinical trial. The trial, the European Carotid Surgery Trial (ECST), was designed to test whether patients who had recent warning signs or symptoms of a stroke benefited from carotid endarterectomy, a surgical procedure to clear plaque from one of the major arteries that carry blood to the head and neck.

This trial was a good candidate for reanalysis because treatment could be harmful as well as beneficial; debris from the artery could break off during surgery and migrate to the brain, causing a stroke. Patients had varying baseline risks for having a stroke if they did not have surgery, and they also had varying baseline risks of suffering harm during surgery. What's more, the factors predicting a patient's baseline risk were different from those predicting his or her risk of stroke during surgery.

The ECST trial showed that if patients had severe or "tight" stenoses (narrowings of the carotid artery that reduced its diameter by 70 percent

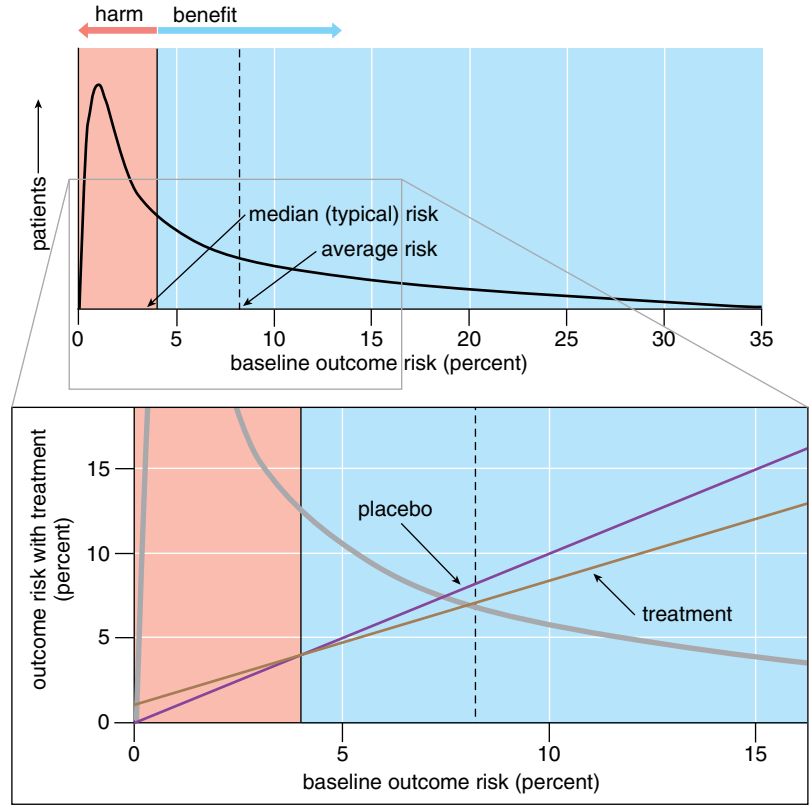


Figure 5. Wide variation of patients' baseline risk (their risk of suffering a bad outcome in the absence of treatment) is one reason trial results don't apply equally to all patients. When the baseline risks are skewed, the average trial results might not even apply to typical patients. The top panel shows a typical skewed risk distribution with many patients at low risk and a few at high risk. The *median* baseline risk in this hypothetical population is less than 4 percent, but the *average* risk is roughly 8 percent because the trial includes a few patients with very high risks that pull up the average. In the inset the purple line shows the expected result for placebo treatment or no treatment. If treatment reduces the risk of the outcome by 25 percent but also carries a risk of harm of 1 percent *(second line)*, the sickest patients would benefit, but those with a baseline risk below 4 percent would actually run a slightly greater risk of being harmed than of being helped. When baseline risk is skewed in this way, a trial may have an overall positive outcome even though most patients are unlikely to benefit.

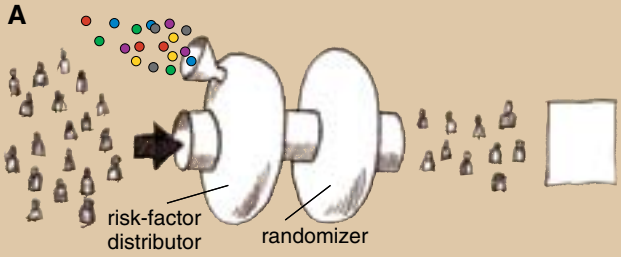
Finding Answers for Vladimir and Estragon

Most physicians are aware that the single-number scores reported for clinical trials can be misleading, and trials are usually analyzed to examine the impact of one or another attribute or risk factor on the effect of an intervention. But these single-factor analyses are much less likely to yield useful insights than is multifactor risk stratification.

The difference in the power of these two types of analyses is demonstrated by a clinical-trial simulation. The virtual patients in this simulation could have any of six different risk factors, each of which increased the risk of a bad outcome by a factor of two. The prevalence of the risk factors varied from 10 percent to 40 percent. Treatment decreased the relative risk of the outcome (a negative event used to measure treatment efficacy) at the five-year mark by 50 percent. But it also led to three bad outcomes per 1,000 patients per year.

On the whole the treatment benefited the virtual population: The treatment-effect, expressed as the relative risk reduction, was 23 percent. The statistical power of the simulated trial *(A)* was also roughly similar to those of many real trials. Given the degree of benefit and the number of patients in the trial, there is a 74 percent chance a single run of the trial would have a statistically significant result.

A

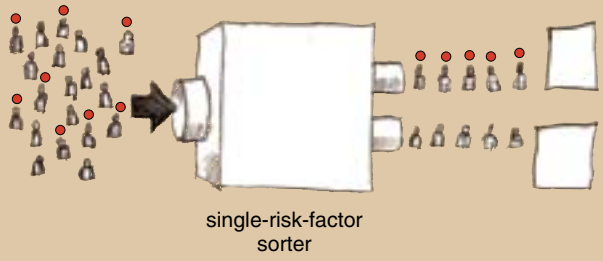


	true control event rate (outcome risk)	true relative risk reduction (treatment-effect)	statistical power
overall results	5.6%	23%	0.74

When the six risk factors were used one by one to divide the patients into two groups *(B)*, the treatment-effect didn't vary substantially among the groups. A single risk factor didn't much change the patients' risk of suffering the outcome, and the presence of other risk factors helped obscure what impact it did have.

Because the differences in the treatment-effect are relatively small, it was unlikely that any one of these single-factor analyses would have statistically significant results. Indeed, even if a risk factor was present in 40 percent of the trial subjects *(bottom row)*, there was only a 19 percent chance that there would be a significant difference between the two groups in the treatment-effect.

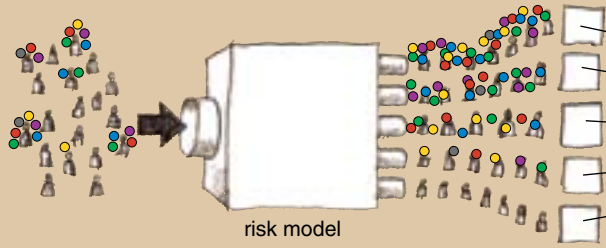
B



risk factor			
#1 present in 10% of subjects	9.5%	33%	0.11
absent in 90% of subjects	5.1%	20%	
#3 present in 25% of subjects	8.5%	30%	0.16
absent in 75% of subjects	4.6%	18%	
#6 present in 40% of subjects	7.8%	29%	0.19
absent in 60% of subjects	4.2%	14%	

In contrast, when a risk index or score was used to divide patients into subgroups *(C)*, we saw tremendous variation both in the outcome risk and the treatment-effect. The lowest-risk group had an outcome risk of 1.5 percent; the highest-risk group had a risk of 23.1 percent. Because of the variation in outcome risk, there was also large variation in the treatment-effect. The treatment-effect reversed sign for the lowest-risk group. Their likelihood of suffering a bad outcome *increased* 49 percent (although this corresponds to just a small absolute increase in risk). The highest-risk group had a 40 percent relative risk reduction with treatment. Given the large variation in treatment-effect, the analysis is quite powerful. The chance of finding a statistically significant difference between risk strata is almost as high as finding a treatment-effect in the overall trial.

C



4–6 risk factors	23.1%	40%	0.71
3 risk factors	11.6%	34%	
2 risk factors	6.3%	17%	
1 risk factor	3.2%	7%	
0 risk factors	1.5%	(49%)	

If the six factors are used to assign each patient a risk score and the patients are then divided into two groups of roughly equal size based on their scores *(D)*, the variation in the outcome risk and the treatment-effect remains large and the statistical power relatively high. Together these simulations demonstrate that more realistic risk analysis is likely to improve the treatment decisions physicians and patients must make.

D

2 or more risk factors	9.0%	31%	0.58
0–1 risk factors	2.6%	(3%)	

or more), they benefited from endarterectomy, which reduced their five-year absolute risk of suffering a stroke by an average 7 percent. According to the overall trial results, *all* symptomatic patients with this risk factor should undergo surgery. Rothwell and Warlow reanalyzed the results for this group of patients.

The scientists derived two models for patient risk (risk of future stroke if untreated and risk of stroke during surgery) from other data. They then used these models to divide the patients with tight stenoses into subgroups and looked at the outcomes in these subgroups. It turned out that among patients with tight stenoses, only 16 percent benefited from surgery. Those who benefited were at relatively high risk of stroke if not treated but at relatively low risk of stroke during surgery. The other 84 percent of the patients had nearly identical outcomes with or without surgery. Again, although the *average* outcome suggested patients benefited, the *typical* patient did not. Reanalysis showed that only one in five patients with tight stenoses was helped by surgery.

Consider, for example, Cesario and Viola. Four days ago, Cesario—who is 76 years old—suddenly, though temporarily, lost control of his right hand and the ability to talk. A cerebral angiogram (an x-ray image of the arteries that supply the brain) showed that his carotid artery was 90 percent blocked, and the plaque had a highly irregular border. Based on this, according to the ECST model, his risk of a stroke over the next five years is over 40 percent.

Viola, on the other hand, is 59 years old. More than three months ago, she experienced transient loss of vision in one eye (suggesting a clot in the vessels supplying her eye rather than her brain). She has had no symptoms since. Her carotid blockage was 70 percent, and the plaque was smooth. Based on Viola's characteristics, her risk of stroke in the next five years is less than 5 percent. Moreover, because Viola is female and has very high blood pressure, her risk of stroke from the surgery is higher than Cesario's. For Cesario the benefits are clear; for Viola, the risks of stroke from the surgery itself would outweigh the benefits.

What Stands in the Way?

Once the benefits of risk-stratified analysis are explained, they seem obvious—so obvious one would think this type of analysis would already be commonplace. Yet risk-stratified analyses are rarely done. In 2001, we reviewed 108 clinical trials reported in four major journals. At that time we found only one trial that used statistical methods similar to those we suggest, and we haven't noticed many since.

Admittedly there are still methodological and practical problems with risk stratification to be ironed out. In a situation where there are factors that affect the outcome risk, factors that

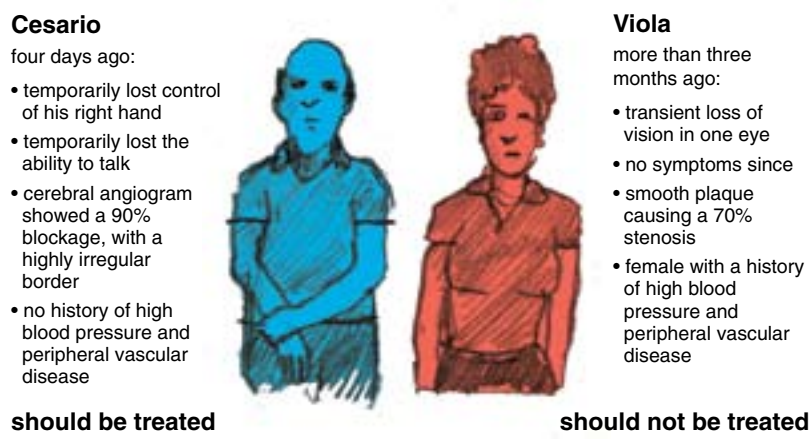


Figure 6. Cesario and Viola, a hypothetical pair of patients who have suffered strokes, illustrate the array of risks facing a physician considering a surgical treatment option. A re-analysis of the ECST clinical trial of carotid endarterectomy (a surgical procedure to clear blocked carotid arteries) relied on two multifactor risk models, one for the baseline risk of stroke without treatment and another for the risk that the patient would suffer a stroke during the surgery itself. Although the original trial resulted in the recommendation that everyone with more than 70 percent blockage of an artery should be sent to surgery, risk-benefit profiling revealed a more complex picture: In fact only one in five of those patients actually stood to benefit from the surgery.

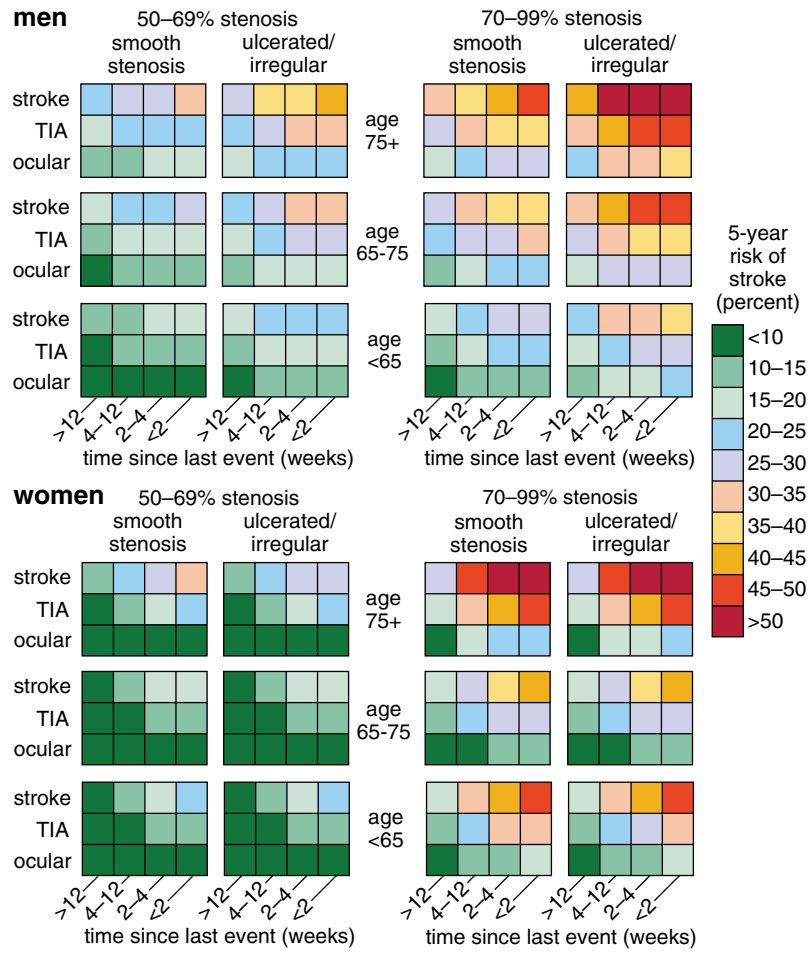


Figure 7. Following the reanalysis of the trial described in Figure 6, investigators suggested that a chart like that shown above might make the results of this kind of analysis more accessible to physicians. The chart presents the five-year risk of a new stroke based on five patient characteristics in addition to gender: time since last stroke, nature of the patient's most severe recent event (a TIA is a transient ischemic attack), age, amount of stenosis and nature of stenosis. (Adapted from Rothwell *et al.* 2005.)

affect treatment-related harm and other factors that affect the responsiveness of patients to therapy, it can be difficult to know how best to combine these different dimensions to appropriately stratify patients. Also, there is never just a single way to describe risk. Different risk models using different variables may be equally valid but place individual patients into different strata, yielding different treatment recommendations depending on which model or score is applied. It should be noted, however, that the ambiguity about what to do for some individual patients currently exists; it is merely obscured by our insistence that the average benefit applies to all.

Additionally, moving risk stratification into the clinic will mean developing a usable set of tools for doctors to predict and communicate risk—and overcoming barriers to their adoption. Such tools might come in the form of charts, such as that shown in Figure 7, or in the form of informatics tools, such as the electrocardiograph-based predictive instruments developed by our coworker Harry Selker and other colleagues at Tufts-New England Medical Center (and used in our GUSTO reanalysis).

Finally, it must be recognized that there are considerable disincentives to taking this approach. What we call "individualized therapy" the pharmaceutical companies that sponsor most drug trials might call "market segmentation." If a trial results in a recommendation that all patients be treated, why look further and perhaps discover that only a subgroup of the patients is really benefiting? Indeed, the only robust risk-stratified analysis we found in our literature survey was done for a trial that showed no benefit overall but demonstrated beneficial effects in higher-risk patients.

Given these impediments, risk stratification, like the clinical trial itself, might not be widely adopted until regulatory agencies require it as part of the drug-approval process. To our knowledge the FDA has linked drug approval to a risk score only once. The PROWESS trial, published in 2001, showed that a new drug, drotrecogin, reduced mortality by 6.1 percent in patients with sepsis, organ failure caused by blood infection. Drotrecogin is a genetically engineered version of a protein normally found in the body that reduces clotting and inflammation. Because the drug is extremely expensive (about $7,000 per patient), the FDA advisory committee required that a risk-stratified analysis be performed on the PROWESS results.

When the patients were stratified according to the APACHE II model (a well-known risk model), it turned out that the half of patients with lower APACHE scores fared no better with the agent, whereas the higher-risk patients benefited much more than the overall result suggested they would. The FDA approved drotrecogin only for these high-risk patients. Perhaps because they believed their overall results more than the risk-stratified results, the makers of drotrecogin then ran a second clinical trial limited to the lower-risk patients. This trial (the ADDRESS trial) confirmed that drotrecogin does not benefit low-risk patients and may instead cause serious complications.

Because risk-based subgroup analyses are so rare, it is impossible to know how often this kind of clinically important variation in benefit goes undetected and leads harmfully to over- or under-treatment. One might say that the conventional approach to reporting overall results of clinical trials consigns us to an impoverished perspective similar to that described in Edwin Abbott's 19th-century science-fiction novella *Flatland*. In Flatland, characters inhabit a two-dimensional plane and perceive objects only if they intersect this plane; the world of three dimensions is unfathomable. In our medical Flatland, all the rich data from a trial is flattened into a single effect; a therapy either works or it doesn't. This binary outcome seems useful, since it conforms well to the binary decisions doctors must make: to treat or not treat. But the treatment decision is easy only because it's fitted to the average patient, not to real individuals. Analyzing and presenting clinical trial results across dimensions of risk can provide us with a more flexible, multidimensional evidence base for treating actual, not average, patients.

For relevant Web links, consult this issue of *American Scientist Online*:

http://www.americanscientist.org/IssueTOC/issue/921

Bibliography

Hayward, R. A., D. M. Kent, S. Vijan and T. P. Hofer. 2005. Reporting clinical trial results to inform providers, payers and consumers: the need to assess benefits and harms for lower vs. higher risk patients. *Health Affairs* 24:1571-1581.

Hayward, R. A., D. M. Kent, S. Vijan and T. P. Hofer. 2006. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. B*MC Medical Research Methodology*. 6:18.

Ioannidis, J. P., and J. Lau. 1997. The impact of high-risk patients on the results of clinical trials. *Journal of Clinical Epidemiology* 50(10):1089-1098.

Ioannidis, J. P., and J. Lau. 1998. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *American Journal of Epidemiology* 148(11):1117–1126.

Kent, D. M., R. A. Hayward, J. L. Griffith *et al*. 2002. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. *American Journal of Medicine* 113:104–111.

Kent, D. M., R. Ruthazer and H. P. Selker. 2003. Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? *Stroke* 34:464–467.

Rothwell, P. M. (editor). Forthcoming. *Treating Individuals: From Randomised Trials and Systematic Reviews to Personalised Medicine in Routine Practice*. London: The Lancet.

Rothwell, P. M., Z. Mehta, S. C. Howard, S. A. Gutnikov and C. P. Warlow. 2005. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet* (Jan 15–21) 365(9455):256-265.