

Limitations of Applying Summary Results of Clinical Trials to Individual Patients

The Need for Risk Stratification

David M. Kent, MD, MS

Rodney A. Hayward, MD

THERE IS GROWING AWARENESS THAT THE RESULTS OF randomized clinical trials might not apply in a straightforward way to individual patients, even those within the trial. Although randomization theoretically ensures the comparability of treatment groups overall, there remain important differences between individuals in each treatment group that can dramatically affect the likelihood of benefiting from or being harmed by a therapy.¹⁻⁴ Averaging effects across such different patients can give misleading results to physicians who care for individual, not average, patients.

The limitations of subgroup analyses—the conventional means for exploring differences in treatment effect based on patient characteristics—are well-appreciated.⁵⁻⁷ Because patients in trials, as in clinical practice, have many attributes that can affect the likelihood of treatment being beneficial or harmful, exploring each of these attributes “one variable at a time” (eg, male vs female, old vs young) risks spurious false-positive subgroup results from chance fluctuations. Furthermore, although patients have simultaneous multiple characteristics that can affect the likelihood of the outcome and the effect of therapy, one-variable-at-a-time comparisons are fundamentally limited because they compare groups that vary only on a single factor, usually resulting in the subgroups being more similar than different.⁸

Two Illustrative Cases

For example, consider the benefits that 2 patients with type 2 diabetes mellitus and mild hypertension (blood pressure of 140/90 mm Hg) might get from intensive blood pressure management. Dick is a 65-year-old smoker with a 15-year history of diabetes, a glycosylated hemoglobin level of 8%, a total cholesterol level of 200 mg/dL (to convert to mmol/L, multiply by 0.0259), and a high-density lipoprotein (HDL) level of 30 mg/dL (to convert to mmol/L, multiply by 0.0259). Jane is a 50-year-old nonsmoker with newly diagnosed diabetes, a glycosylated hemoglobin level of 7%, a total cholesterol level of 180 mg/dL, and an HDL level of 55 mg/dL. Based on these characteristics, the 5-year risk of cardiovascular death for Dick is more than 20% and for Jane is less than 1%.⁹ If intensive blood pressure control decreased cardio-

vascular mortality by 10%, that would be important for Dick, but less so for Jane. Indeed, if intensive treatment carried even a small risk of serious harm from polypharmacy, the risks for Jane could easily outweigh the benefits. However, in a clinical trial that tested tight blood pressure control in persons with diabetes, Dick and Jane would be lumped together and the small net risks to patients like Jane would be obscured within the overall average results, overwhelmed by the benefits to patients like Dick.

As another example, consider 2 patients presenting with acute myocardial infarction (AMI). Bud, aged 51 years and otherwise healthy, presents with a small inferior wall AMI, a blood pressure of 164/90 mm Hg, and a stable heart rate. Lou, a 78-year-old patient with diabetes, presents with tachycardia, hypotension (90/60 mm Hg), and a large anterior wall AMI. The acute mortality risk is around 2% for Bud, but 20% for Lou.¹⁰ If thrombolysis yields a 25% relative reduction in mortality, a modest benefit is expected for low-risk Bud, but a substantial benefit is expected for high-risk Lou. With a 1% risk of thrombolytic-related intracranial hemorrhage (and higher in patients with high blood pressure like Bud), treatment decisions should be different for these 2 patients. It would matter little to Bud that the treatment was found to be beneficial overall in the randomized controlled trial when his personal risk of thrombolytic-related harm is more than twice that of his chance of thrombolytic-related benefit.

Clinically Important Differences in Treatment Effect, Which Are Only Detectable Using Risk-Based Analyses, Are Likely to Be Common

Risk-based analysis is a statistically powerful tool for uncovering variation of treatment effect in patients like those described above.^{4,8,11} However, because such risk-stratified analyses are rarely performed, it is not known how frequently substantive differences in net treatment effect remain undetected in clinical trials. In fact, it is not even known how often the summary results of a clinical trial apply to

Author Affiliations: Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, Boston, Massachusetts (Dr Kent); and Veterans Affairs Ann Arbor Health Services Research and Development Service Center of Excellence and Department of Internal Medicine, University of Michigan, Ann Arbor (Dr Hayward).

Corresponding Author: David M. Kent, MD, MS, Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, 750 Washington St, Box 63, Boston, MA 02111 (dkent1@tufts-nemc.org).

most of the patients in the trial. Counterintuitively, it does not take extreme assumptions to generate conditions where the summary results of a trial do not even apply to the typical patient in the trial. Such conditions are likely to be common and clinical trial results should routinely assess for this by using multivariate risk-stratified analysis.^{4,8,11}

Substantial Variation of Individual Baseline Risk Within a Clinical Trial Is Common, and Often Extreme, Almost Ensuring Marked Variation of the Absolute Treatment Benefit Across Individuals. Large variation in outcome risk between patients in clinical trials appears to be common.¹² For AMI, if patients are stratified using a risk index based on easily obtainable pretreatment variables, the mortality rate in the quartile of patients at highest baseline risk can be more than 10-fold that in the lowest-risk quartile.^{13,14} For human immunodeficiency virus trials, outcome rates have been found to be as much as 46 times higher in the highest risk compared with the lowest risk quartile of study participants,¹² and in a pooled sample of trials examining chronic kidney disease progression, this ratio was about 70.¹⁵ When outcome rates differ this much between large subgroups of patients, the degree of benefit and the trade-offs between benefits and harms will necessarily differ substantially across

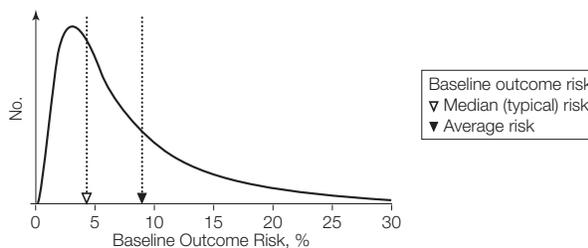
patients even when the treatment has similar relative effectiveness for all.

Baseline Risk Is Typically Highly Skewed, Ensuring That the Average Risk (and Average Treatment Effect) Observed in the Summary Results of the Trial Will Be Different From That in the Typical Patient. If baseline risk were normally distributed, the mean benefit for a treatment observed in the trial's summary results would at least reflect the treatment effect found in the typical patient enrolled in the trial. But this is frequently not the case. Often, most trial outcomes occur in a relatively small number of high-risk patients, while most patients are at much lower than average risk.¹⁶ This is because risk variables often cluster in patients. Older patients typically have more risk factors (apart from age) than younger patients and patients with diabetes typically have more risk factors, such as hypertension and dyslipidemias, than patients without diabetes.² Also, since clinical trial outcomes are generally infrequent, a "floor effect" prevents a normal bell-shaped risk distribution, because patients can and will have risks much higher than average but cannot have risks less than 0. Thus, it is common, as for AMI trials, that there are many more low-risk patients (like Bud) than high-risk patients (like Lou). The overall trial results,

Figure. Population Distribution of Baseline Outcome Risk, Outcome Risk With Treatment, and Relative Risk Reduction

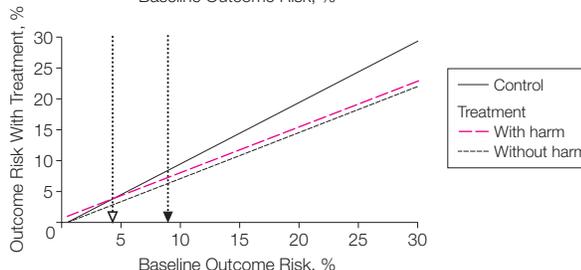
A Population Distribution of Baseline Outcome Risk

Patients enrolled in clinical trials often have greatly different baseline risks for the outcome of interest. The risk distribution is often skewed; a relatively small group of high-risk patients with multiple risk factors account for a large number of the outcomes and the mean risk might be considerably higher than the risk in the typical (median) patient.



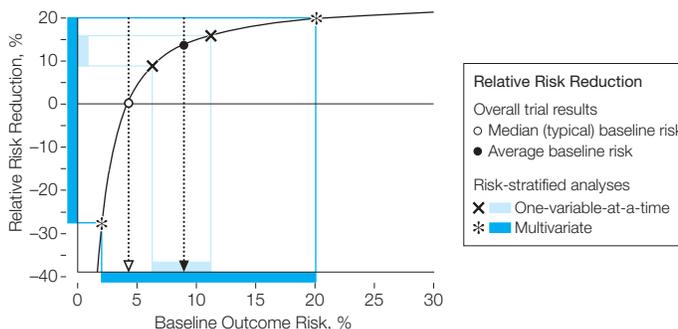
B Outcome Risk With Treatment

A constant relative risk reduction (25% in this case) leads to increasing benefits as baseline risk increases; treatment and control outcome rates progressively diverge at higher baseline risks. When a therapy is associated with even a small amount of treatment-related harm, low-risk patients are unlikely to benefit at all. When the treatment-related risk of harm is 1%, patients with baseline risks lower than 4% have net harm from the therapy. The average baseline risk of the enrolled patients will determine whether the trial's summary results are positive overall. But the overall results may not reflect the trade-offs between the risks and benefits of many individual patients in the trial.



C Relative Risk Reduction

There is considerable variation in relative risk reduction given the assumptions of risks and benefits shown in B. The overall trial results (average baseline risk) indicate a 12.5% relative risk reduction but the typical patient (median baseline risk) does not benefit at all. One-variable-at-a-time subgroup analyses typically compare groups of patients that do not differ dramatically from the average risk (a 2-fold difference in risk), because the treatment effect differences may not be statistically significant, which can misleadingly imply a consistent treatment effect. Using multivariate risk indices compares patients across a broader range of baseline risks, exposing larger differences in the relative treatment effect, which are often clinically and statistically significant.



however, will reflect the arithmetic mean, which is unduly influenced by the relatively small high-risk group. In this type of distribution, the typical (median-risk) patient is at much lower risk than the average (FIGURE, A).

When There Is Substantial Variation in Baseline Risk Across Study Patients, the Presence of Even a Small Degree of Treatment-Related Harm Ensures Variation in the Net Relative Risk Reduction (Unless the Risk of Treatment-Related Harm Is Highly Correlated With Outcome Risk). This effect is most apparent when the treatment-related harm causes primary outcome events (Figure, B and C). There are many such examples. Thrombolytics for AMI can both prevent and cause death, carotid endarterectomy can both prevent and cause stroke, and blood pressure medicines can both prevent and cause cardiac events. However, even when treatments do not adversely affect the main outcome measure of a trial, treatments are virtually always associated with some risk (known or unknown) of adverse effects. Risk-stratified analyses that identify very low-risk patients who get negligible treatment-related benefit are informative even when such patients get the same relative risk reduction as higher-risk patients, because clinicians can at least factor in the potential harms of recommending additional treatments.

Conventional Subgroup Analyses Are Typically Inadequate to Detect These Large and Clinically Important Differences in Treatment Effect Among Patients When Multiple Factors Determine Risk. A series of simulations⁸ showed that even when (1) a substantial number of study participants receive net harm from a therapy that is beneficial overall, (2) the trial is well-powered, and (3) subgroup analyses are based on common and important risk variables, one-variable-at-a-time subgroup analyses are highly unlikely to detect differences in the treatment effect between subgroups (ie, low statistical power) and will lead to the misleading conclusion that the treatment effect is consistent across patients. This is because a single risk factor will rarely dramatically affect overall patient risk, and thus treatment effect differences are often small in one-variable-at-a-time analyses (Figure, C).

Risk-Stratified Analyses Greatly Increase the Power of Detecting These Differences in Treatment Effect. Because of the limitations of conventional subgroup analysis, a common assumption is that there is no alternative. However, a single analysis, stratifying patients by using a multivariate risk model, frequently (although not always) will be sufficiently powered to detect important treatment differences whenever substantial heterogeneity in risk exists within the study population.⁸ This can generally be accomplished with even a moderately predictive multivariate prediction tool,⁸ because a multivariate risk-stratified analysis compares the treatment effect in patients across a fuller spectrum of baseline risk (often ranging in risk by 10 to 30 fold) than a conventional one-variable-at-a-time subgroup analysis (few independent risk factors increase risk by more than 2 fold).

Treatment effects are much more likely to differ in patients with greatly different outcome rates (Figure, C). Multiple and varied examples have demonstrated that multivariate risk-stratified analyses can be clinically informative, revealing large subgroups with results that differ substantially from the overall summary.^{13,17-22}

The key feature of a risk-stratified analysis is that several patient attributes (or risk factors) are combined into a score that describes a single dimension of risk along which treatment effect is likely to vary (almost always on the absolute risk scale, and potentially on the relative risk scale as well). Thus, analyses of this type minimize the major shortcomings of one-variable-at-a-time subgroup analyses (multiple comparisons resulting in false-positive findings and poor statistical power resulting in false-negative findings).

Other methodological issues certainly need to be addressed. Identifying and combining factors that influence additional dimensions of risk, aside from outcome risk, may also be important in determining treatment benefit. These include the risk of treatment-related harm,^{13,17,21} the risk of competing outcomes unresponsive to therapy, and factors that directly modify treatment effect.⁸ Also, validated multivariate risk indices do not exist for all conditions and outcomes, and those that do exist often show substantially worse performance when tested by independent researchers.

Despite these and other caveats,^{4,8} currently there are methods to produce better scientific evidence about the risks and benefits of treatments for patients; however, recommendations continue to be based on crude averages. An important barrier to more individualized therapies is not a paucity of good predictors of outcomes and effects, which are already abundant even in the pregenomic era, but the lack of a consistent analytic approach that informs how an individual patient's multiple attributes combine to affect the fundamental determinants of the desirability of treating that patient—the individual's risk of bad outcomes in the absence of the treatment vs the individual's risk of bad outcomes if treated. Multivariate risk-stratified analyses based on easily obtainable clinical variables are often informative and frequently feasible, but rarely performed. Making such analyses standard should be seriously considered.

Financial Disclosures: Dr Kent reports receiving research support from Pfizer. No other disclosures were reported.

Funding/Support: This work was supported in part by a Career Development Award (K23 NS44929-01) from the National Institute for Neurological Disorders and Stroke and by grant QUERI DIB 98-001 from the Veterans Affairs Health Services Research and Development Service. Additional support was provided by grant P60 DK-20572 from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health.

Role of the Sponsors: The funding agencies had no role in the preparation, review, or approval of the manuscript.

REFERENCES

1. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet*. 1995;345(8965):1616-1619.
2. Vijan S, Kent DM, Hayward RA. Are randomized controlled trials sufficient evidence to guide clinical practice in type 2 (non-insulin dependent) diabetes mellitus? *Diabetologia*. 2000;43(1):125-130.

3. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment-effects, and the trouble with averages. *Milbank Q*. 2004;82(4):661-687.
4. Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers, and consumers. *Health Aff (Millwood)*. 2005;24(6):1571-1581.
5. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069.
6. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116(1):78-84.
7. Rothwell PM. Treating individuals 2: subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*. 2005;365(9454):176-186.
8. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol*. 2006;6:18.
9. UKDPS Risk Engine. The Oxford Centre for Diabetes, Endocrinology & Metabolism. Diabetes Trial Unit Web site. <http://www.dtu.ox.ac.uk/index.php?maindoc=/riskengine/>. Accessed February 5, 2007.
10. Selker HP, Griffith JL, Beshansky JR, et al. Patient-specific predictions of outcomes in myocardial infarction for real-time emergency use: a thrombolytic predictive instrument. *Ann Intern Med*. 1997;127(7):538-548.
11. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet*. 2005;365(9455):256-265.
12. Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *Am J Epidemiol*. 1998;148(11):1117-1126.
13. Kent DM, Hayward RA, Griffith JL, et al. 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. *Am J Med*. 2002;113(2):104-111.
14. Kent DM, Ruthazer R, Griffith JL, et al. Comparison of mortality benefit of immediate thrombolytic therapy versus delayed primary angioplasty. *Am J Cardiol*. 2007;99(10):1384-1388.
15. Kent DM, Jafar TH, Hayward RA, et al. Progression risk, urinary protein excretion and the treatment-effects of angiotensin converting enzyme inhibitors in non-diabetic nephropathy. *J Am Soc Nephrol*. 2007;18(6):1959-1965.
16. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. *J Clin Epidemiol*. 1997;50(10):1089-1098.
17. Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. *Lancet*. 1999;353(9170):2105-2110.
18. Food and Drug Administration. Xigris: drotrecogin alfa (activated): PV 3420 AMP. Indianapolis, IN: Eli Lilly & Co; 2001. <http://www.fda.gov/cder/foi/label/2001/droteli112101LB.pdf>. Accessibility verified August 9, 2007.
19. Antman EM, Cohen M, Bernink PJ, et al. The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *JAMA*. 2000;284(7):835-842.
20. Morrow DA, Antman EM, Snapinn SM, et al. An integrated clinical approach to predicting the benefit of tirofiban in non-ST-elevation acute coronary syndromes: application of the TIMI Risk Score for UA/NSTEMI in PRISM-PLUS. *Eur Heart J*. 2002;23(3):223-229.
21. Kent DM, Ruthazer R, Selker HP. Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? *Stroke*. 2003;34(2):464-467.
22. Thune JJ, Hoefsten DE, Lindholm MG, et al. Simple risk stratification at admission to identify patients with reduced mortality from primary angioplasty. *Circulation*. 2005;112(13):2017-2021.

EDITORIALS

Editorials represent the opinions of the authors and JAMA and not those of the American Medical Association.

B Vitamins for the Prevention of Vascular Disease

Insufficient Evidence to Justify Treatment

Colin Baigent, BM BCH, FRCP

Robert Clarke, MD, FRCP

INDIVIDUALS WITH HOMOCYSTEINURIA, A RARE GENETIC DISORDER associated with markedly elevated plasma homocysteine levels (ie, 100-300 $\mu\text{mol/L}$ [to convert to mg/dL , divide by 7.397]—at least 10 times higher than the general population), experience rapidly progressive atherosclerosis and associated thromboembolic events in early adulthood. This observation prompted the “homocysteine hypothesis” that moderately elevated homocysteine levels might be of causal relevance to cardiovascular disease in the general population.¹

Many observational epidemiological studies have reported that cases with coronary heart disease (CHD) or stroke have higher homocysteine levels compared with age and sex-matched controls. In 1995, a meta-analysis of observa-

tional studies, involving a total of 2297 CHD cases, indicated that a 5 $\mu\text{mol/L}$ higher homocysteine level was associated with about a 70% increase in the risk of CHD.² These studies were mainly retrospective in design, however, and were unable to exclude the possibility that the atherosclerotic disease process might itself have increased homocysteine levels (reverse causality bias). In 1998, an updated meta-analysis of both retrospective and prospective (or nested case-control) studies reported weaker associations in prospective studies than in retrospective studies; a 5 $\mu\text{mol/L}$ higher homocysteine level was associated with only a 30% increased risk of CHD within the prospective studies.³

Because homocysteine levels are closely correlated with established risk factors for vascular disease, meta-analyses of published results of observational studies are unable to

Author Affiliations: Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, Oxford, England.

Corresponding Author: Colin Baigent, BM BCH, FRCP, Richard Doll Building, Old Campus Road, Roosevelt Drive, Oxford OX3 7LF, England (colin.baigent@ctu.ox.ac.uk).

See also p 1163.