ABSTRACT
This review addresses decision-making underlying the frequent failure to confirm early-phase positive trial results and how to prioritize which early agents to transition to late phase. While unexpected toxicity is sometimes responsible for late-phase failures, lack of efficacy is also frequently found. In stroke as in other conditions, early trials often demonstrate imbalances in factors influencing outcome. Other issues complicate early trial analysis, including unequally distributed noise inherent in outcome measures and variations in natural history among studies. We contend that statistical approaches to correct for imbalances and noise, while likely valid for homogeneous conditions, appear unable to accommodate disease complexity and have failed to correctly identify effective agents. While blinding and randomization are important to reduce selection bias, these methods appear insufficient to insure valid conclusions. We found potential sources of analytical errors in nearly 90% of a sample of early stroke trials. To address these issues, we recommend changes in early-phase analysis and reporting: (1) restrict use of statistical correction to studies where the underlying assumptions are validated, (2) select dichotomous over continuous outcomes for small samples, (3) consider pooled samples to model natural history to detect early therapeutic signals and increase the likelihood of replication in larger samples, (4) report subgroup baseline conditions, (5) consider post hoc methods to restrict analysis to subjects with an appropriate match, and (6) increase the strength of effect threshold given these cumulative sources of noise and potential errors. More attention to these issues should lead to better decision-making regarding selection of agents to proceed to pivotal trials.
We contend that reasons for failure to replicate apparently positive early trial results include imbalances in key factors related to outcome and nonrandom distribution of errors in assessments and outcome measures. Importantly, many early trials applied statistical methods to correct or adjust for imbalances. A closer examination of the assumptions underlying these methods suggests they are not appropriate for this purpose. The inappropriate use of statistical correction is a fixable cause of potentially false-positive and negative early-phase conclusions. We offer solutions to increase rigor in analysis and reporting so the larger community can make a more considered opinion as to the value of further investment in a therapy.

**ADJUSTMENT FOR IMBALANCES** We suggest that the application of statistical methods such as multivariate regression and covariate adjustment intended to adjust or correct for baseline imbalances often led to misleadingly positive early results. While these analytical methods may be appropriate to homogenous conditions, there has been little attention to whether the cases in which these methods are used meet the assumptions necessary for the proper application of these methods. Imbalances are a particular problem when subgroups are explored post hoc because parsing subjects into smaller samples increases the likelihood that imbalances in baseline factors will be accentuated.

The basis of our criticism of trial analytical methodology is the complexity of the relationships between baseline factors. Despite randomization, because most early trials are relatively small, they often generate treatment arms that are imbalanced in prognostic factors such as stroke severity, usually measured by the NIH Stroke Scale (NIHSS). Initial stroke severity remains the most robust predictor of outcomes. Importantly, the relationship between NIHSS and functional outcome is not linear (figure 1). Imbalances have been seen in many reasonably sized stroke trials, as shown in table e-1 on the Neurology® Web site at Neurology.org. While the differences may appear to be small, even minor differences that are not statistically significant can influence outcome, especially in the middle ranges of stroke severity, where the relationship between changes in baseline stroke severity and outcome is steep (figure 1).

Imbalances are important to consider because most trial outcomes are based on the percentage of subjects who achieve certain endpoints, which could be influenced by the distribution of severity within the different arms. If a treatment is neutral but, by random chance, the placebo arm consisted of subjects with a poorer prognosis, then a treatment may appear to be effective only because the placebo-treated patients had a worse natural history. Similar error may occur if by chance the study’s placebo population was not representative of a broader population that would be expected if the study was performed in a larger number of sites. These errors are less likely to occur in a large phase 3 trial, although even those sometimes show imbalances (table e-1). The converse may result in a false-negative result if the treatment arm imbalances resulted in a worse outcome.

There have been many different approaches to deal with imbalances. Most trialists address imbalances by post hoc statistical methods such as multivariate analysis, covariate correction, or the widely used stratified Cochran-Mantel-Haenszel (CMH) test. Our concern with these techniques is that for proper application, there must be linear relationships between the factors and the outcome and the coefficients of the relationship between the factors and outcome are assumed to be the same for both arms. The CMH test further assumes that the odds ratios being calculated for each strata are the same. It is uncommon that these assumptions are verified and reported before these methods are used.
applied. Examination of the frequency distribution curves of NIHSS and age, for example in the National Institute of Neurological Disorders and Stroke recombinant tissue plasminogen activator (rtPA) trial, demonstrates very complex relationships that are neither linear nor Gaussian (figure 2). Even in this well-executed trial, a distribution of lower NIHSS is evident in the treatment arm and excess severe NIHSS is seen in the placebo arm. Nevertheless, our and other reanalysis of this trial indicated benefit persisted, reflecting the robustness of the treatment effect that other therapies apparently do not possess.

An analysis of early and late trial results suggests that these correction methods are not valid in other stroke trials. An example of application of the CMH test with stratification and correction for several factors is the early-phase citicoline trial, with a positive result reported after adjustment. These positive results were used as the basis for the follow-on International Citicoline Trial on Acute Stroke (ICTUS) trial, which was subsequently negative. It is conceivable that the CMH test did not reliably adjust for baseline imbalances in the earlier trial. This possibility is supported since, unlike the corrected results, uncorrected outcomes of the first trial were negative, the same result obtained in the larger follow-up. Similarly, statistical adjustment failed to predict a positive outcome in the Albumin in Acute Stroke (ALIAS) trial, a study that was performed in 2 parts. Part 1 employed multivariate adjustment and suggested a trend of efficacy in the subgroup that received tissue plasminogen activator. However, part 2 of phase 3 ALIAS trial ultimately failed to show efficacy in any group.

**MODELING HISTORICAL OUTCOMES CAN SUBSTITUTE FOR POST HOC CORRECTION**

Given that any given trial may have an imbalance in one or more factors, modeling outcomes of a large pooled sample may provide one approach to assessing the importance of an early finding by comparing predicted results at the trial’s own baseline values. Comparison with an outcome model can also suggest whether the study’s control population is representative of a broader sample that would be expected if the trial was expanded to include multiple sites. Several groups have modeled outcome from acute stroke, with most finding that baseline stroke severity, usually based on the NIHSS or a variant, is the strongest predictor of functional outcome and mortality, with additional contribution from other factors including age. We modeled outcomes based on the placebo arms of stroke randomized controlled trials (RCTs) to generate a surrogate control function.
that incorporates 47 trials and represents more than 11,000 subjects (figure 3, middle surface). Generating this model and multidimensional statistical intervals around this function can be used to indicate whether a study’s control arm is representative of a larger population and whether the treatment arm is meaningfully different from a pooled control sample (figure 3). Despite the large number of trials, there is still variation around the function. Potential sources of additional variability include both tangible and intangible factors including variation in the distribution of stroke etiologies with different propensities to bleed or progress, blood glucose, race, sex, genetic factors, and psychosocial factors, each of which can influence outcome and are unlikely to be balanced in a small trial. These sources of variability probably raise the necessary threshold for a minimum strength

Figure 3 Good outcome (modified Rankin Scale 0-2) proportions compared to baseline median NIH Stroke Scale and mean age with superimposed results from 4 different medication trials at their respective baseline conditions

The middle surface is the model function derived from 28 randomized controlled trial control arm outcomes. Surfaces on either side represent the ±95% prediction intervals. (A) Parts 1 and 2 of the National Institute of Neurological Disorders and Stroke (NINDS) treatment and control arms. Both control arms are near the middle surface, indicating the results are representative of the pooled control sample. Both recombinant tissue plasminogen activator (rt-PA) arms are above the 95% interval, indicating they are significantly better at the \( p < 0.05 \) level. (B) Stroke–Acute Ischemic NXY Treatment (SAINT) I and SAINT II treatment and control arms. The control and treatment arms for both trials are essentially superimposable and all lie on the control surface, indicating no signal of efficacy was apparent for this outcome measure in either trial. (C) Treatment and control arms of phase 2 trial of abiciximab (Abiciximab in Emergency Treatment of Stroke Trial [AbESTT]) and the phase 3 trial of abiciximab (AbESTT II and AbESTT IIc) treatment and control arms. There was no signal of efficacy in either early or late since all arms lie essentially on the control surface. (D) Treatment and control arms of phase 3 trial of citicoline completed in 2001 and International Citicoline Trial on Acute Stroke (ICTUS) trial completed in 2012. There was a higher proportion of subjects in the treatment arm of the initial trial that achieved a modified Rankin Scale (mRS) 0-2, but it can be appreciated that this was because of less severe baseline NIH Stroke Scale (NIHSS) rather than a treatment effect since both arms lie on the control surface although in different regions. The follow-up larger ICTUS trial had minimal imbalance between treatment and control arms, but both lie between the control surface and the ±95% interval, indicating overall worse outcomes than in 2001 and no signal of treatment effect. (A–C) Reproduced with permission.
of effect that could be reasonably expected to be reproducible in a larger population.

Examples of early outcomes compared to later phase trials superimposed on a model of the relationship between baseline NIHSS, age, and the proportion that achieved a modified Rankin Scale (mRS) of 0-2 are seen in figure 3. The middle surface represents a nonlinear fit to the median NIHSS and mean age of the control arms from consecutive stroke RCTs and the z axis is the percentage that achieved an mRS 0-2 at 90 days follow-up. The surfaces above and below the middle surface are the \( p = 0.05 \) (±95%) prediction intervals so that a point above or below this surface would predict a positive or negative outcome in a larger sample. In figure 3A, National Institute of Neurological Disorders and Stroke trial\(^{29}\) part 1 and part 2 are plotted. The control arms of both parts are adjacent to the outcome function at their respective baseline NIHSS and age, indicating the control outcomes remain representative of the outcomes expected in more recent trials at comparable baseline NIHSS and age. Importantly, both treatment arm results are above the upper \( p = 0.05 \) surface, indicating that the initial part 1 would have predicted a positive part 2 had the study been designed as such,\(^6\) without any need for statistical correction despite clear imbalances at baseline seen in part 2.

By contrast, the treatment arm results of both the initial and larger follow-up Stroke–Acute Ischemic NXY Treatment (SAINT) I\(^{15}\) and II\(^{36}\) and the Abciximab in Emergency Treatment of Stroke Trial\(^{37,38}\) are essentially no different from expected outcomes, indicating that both the early- and the late-stage trials were essentially negative. This result suggests that any early positive signal was most likely due to imbalances or random noise between the control and the treatment arm. The phase 3 citicoline trial\(^{23}\) completed in 2001 and the recently completed ICTUS\(^{3}\) follow-up trial outcomes are plotted in figure 3D. The outcomes of both the treatment and control arm of the earlier trial are near the model’s control arm surface. An imbalance between the treatment and control arm results of the earlier phase 3 citicoline\(^{23}\) trial are evident, with a more severe NIHSS in the placebo arm. It is conceivable that better outcomes in the less severely affected treatment arm subjects were erroneously interpreted as a drug effect and that statistical correction was unable to accommodate this imbalance. Note that the follow-up trial results with a larger sample size showed no difference in outcomes between placebo and control and actually worse outcome at the same baseline NIHSS and age.

POST HOC MATCHING AS AN APPROACH TO IMBALANCES Many investigators have recognized the role that imbalances play in influencing outcomes. Post hoc matching has been suggested as an alternative approach, but as with nearly all of these issues, there remains disagreement as to the best approach. The propensity score is probably the most widely used balancing approach. It was originally developed to handle large nonrandomized observational studies where there is inherent bias in assignment to the sole treatment arm. Rosenbaum and Rubin\(^{39,40}\) developed this method with the hope that if large enough datasets were used and with judicious selection of appropriate covariates, effect of bias can be nullified.\(^{39,40}\) However, there are critical assumptions that are required for valid use of the propensity score, including the following: (1) The groups have to be large. (2) The 2 groups should have overlapping distributions with respect to covariates. (3) These distributions should be Gaussian.\(^{40-42}\) It is unlikely the typical modest-sized stroke trial achieves these assumptions. While nonlinear methods for propensity score have been proposed,\(^{43}\) it is not clear that they can capture the range of nonlinear distributions expected in the routine stroke clinical trial.

A method with fewer assumptions involves matching subjects in Euclidean space based on prespecified factors of the investigator’s choice.\(^{21,44}\) This method simply simultaneously matches quantifiable factors in multiple dimensions and can identify potential positive signals in early trials when compared to a larger historical dataset.\(^{21,45}\) We employed a variation of Euclidean matching to weight factors based on their numerical range,\(^{21,44}\) so that, for example, glucose levels do not overcome the effect of the more narrow range of NIHSS.

Matching the National Institute of Neurological Disorders and Stroke\(^{46}\) placebo arm patients (\( n = 312 \)) with rtPA-treated patients (\( n = 312 \)) by propensity score matching and Euclidean matching showed much larger distances between points (e.g., city-block distances\(^{46}\) for propensity score (table 1). A city-block distance of zero between 2 matched points indicates a perfect match. Table 1 shows that, while overall balance was improved by both methods, city-block distances are lower for the Euclidean nearest-neighbor method (\( p < 0.001; t \) test), indicating that the Euclidean method identified subjects closer to each other yielding more similar subjects, thereby reducing subject heterogeneity as a source of noise. One downside of outlier elimination is the reduction in overall statistical power. However, a more sophisticated use of power analysis includes a term for population variance; an improvement in homogeneity can potentially counter the effect of a reduction in power due to smaller sample size.\(^{41}\)

CONTRIBUTION OF NOISE IN OUTCOME MEASURES Many conditions employ subjective outcome scales involving levels of impairment or
disability, including Barthel Index, Glasgow Outcome Scale (GOS), and mRS.

The most commonly used outcome measure in stroke is the mRS, which provides a range of outcomes from no disability to death. The National Institute of Neurological Disorders and Stroke rtPA trial used an excellent outcome of mRS 0-1 as one of the primary outcome measures. If a range of outcomes can be expected, variations on this dichotomization method have been proposed. These include a stratified outcome statistic or responder analysis and the full-scale ordinal or shift analysis that uses variations on the entire range of scores. The advantage cited, particularly for the shift analysis, is that more information is communicated by using the whole scale and has been used as a justification for smaller sample size, such as in the truncated third International Stroke Trial (IST-3). While these methods are well-intended to communicate more information regarding a range of outcomes, they neglect to consider the issue that noise or errors in the assessment of a subject’s mRS is not uniformly distributed across

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comparison of propensity score and Euclidean matching for post hoc correction of imbalances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-match</td>
</tr>
<tr>
<td></td>
<td>rtPA (n = 312)</td>
</tr>
<tr>
<td>Median NIHSS</td>
<td>14</td>
</tr>
<tr>
<td>NIHSS, mean ± SD</td>
<td>14.4 ± 7.5</td>
</tr>
<tr>
<td>Age, y, mean ± SD</td>
<td>68.0 ± 11.3</td>
</tr>
<tr>
<td>Glucose, mean ± SD</td>
<td>149 ± 70.7</td>
</tr>
<tr>
<td>Mean city block distance ± SD</td>
<td>17.7 ± 12.45</td>
</tr>
</tbody>
</table>

Abbreviations: NIHSS = NIH Stroke Scale; rtPA = recombinant tissue plasminogen activator.
Baseline factors (NIHSS, age, and glucose) pre-match and post-Euclidean and propensity score match from the National Institute of Neurological Disorders and Stroke trial. Distributions of city-block distances demonstrate much closer matches for Euclidean matching than propensity score, indicating more similar subjects and consequently less potential contribution to errors in assessing group differences. Reproduced with permission from Mandava et al.45

Figure 4
Box plots of error rates with the ordinal scale (shift analysis) and a series of dichotomizations of the modified Rankin Scale

The first 2 columns are 2 variants of the ordinal or shift analysis used in outcome studies. Columns 3-6 are dichotomizations at increasing modified Rankin Scale (mRS) thresholds. Error rates were lower for all dichotomizations vis-à-vis ordinal scales. Error rates in outcome classification increases the required sample size. Figure reproduced under creative commons license.
the range.49 This asymmetric noise distribution exists regardless of the extent of training or interrater agreement.50 There is little error in determining 6 (death), a little more for 0 (no disability), and considerably more in distinguishing the middle ranges. This noise appears to be an intrinsic feature of the subjectivity involved with disability determination.49 The mid-range of GOS, which is predominantly used in traumatic brain injury trials, also has low interrater reliability.11 When this noise is taken into account, it greatly increases overall errors in misclassification vs if a dichotomous mRS was used (e.g., 0-2; see figure 4).52 and results in the need for a larger sample size. When applied to the SAINT I trial’s actual data, consideration of the noise factor with the shift, we calculated the need of nearly 25% larger sample size than were recruited in the actual study.15,52 Small sample size makes it more likely that this noise is nonrandomly distributed between treatment arms, and could explain why the primary positive signal from SAINT I was in the shift,15 but the larger SAINT II did not confirm this finding,56 while the primary dichotomous outcome was negative in both. While additional information transfer is an advantage of ordinal analysis,53 reducing sample size does not seem to be justified. We have provided a set of programs to incorporate outcome measure noise into sample size calculations.54

**POTENTIAL SOURCES OF ERRORS ARE FREQUENT IN EARLY-PHASE TRIALS** Through the above analysis, we identified 6 potential sources of errors in analysis of early-phase trials that could have contributed to misleading evidence for efficacy. These factors are as follows: (1) imbalances in baseline factors that favored better outcomes in the treatment arm, (2) use of inappropriate statistical adjustments to correct these imbalances, (3) use of outcomes measurements that introduce noise and may not be randomly distributed in a smaller early-phase trial, (4) basing further study on a positive subgroup without indicating that the subgroups’ treatment and control arms were balanced, (5) a study’s control arm that is not representative of a typical population, and (6) a futile therapy when compared to a larger historical population. We summarize these potential sources of errors in table 2.

To assess the potential contribution of these factors to the decision to progress through trial phases, we reviewed the clinical trial literature to identify all phase 3 ischemic stroke RCTs that were preceded by early-phase trials and had sufficient information on the early-phase trials to assess for baseline characteristics. We gleaned from these studies the authors’ reasons for suggesting progression to later phases. We identified 16 such agents with 32 preliminary phase trials. At least 1 issue was found in 15 agents and 28 of 32 preliminary trials. The exception is Tirilizad, where it appears that convincing preclinical data prompted progress through phases since there was no early positive signal in its 2 early-phase trials. Table 2 summarizes these results. Imbalances with or without adjustments were common. Comparison with the pooled model would have identified several agents as futile or unrepresentative control arms, although many trials did not provide sufficient information to assess their likely success. Table e-2 provides details on each of these failed agents and their earlier trials including the factors identified and the author’s original interpretation of the trial.

**INNOVATIVE STUDY DESIGN MAY INTRODUCE ERRORS IN EARLY PHASE** New approaches to enhance recruitment and speed the trial process may also be susceptible to these problems, especially when applied to early phases. Adaptive designs include minimization routines that identify key prognostic factors to guide treatment assignments.54,55 When used in large trials such as IST-3, overall balance is achievable, but correction methods were still necessary for subgroup analysis.49 For small trials, it is not clear that improvement in balance can be achieved without a risk of selection bias.49 Another approach is the prospective, randomized, open, blinded endpoint evaluation design used in IST-3,48 the recently completed Multicenter Randomized Clinical Trial of Endovascular Treatment for Acute Ischemic Stroke in the Netherlands (MR CLEAN),56 and secondary stroke prevention trials such as Randomized Evaluation of Long-Term Anticoagulation Therapy (RE-LY).57 We suggest caution in using any method that does not include patient blinding and randomization given the

---

**Table 2** Frequency of potential errors in the lead up to 16 negative phase 3 trials

<table>
<thead>
<tr>
<th>Factor identified</th>
<th>Number founda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of early-phase trials with an identified potential source of error</td>
<td>28/32</td>
</tr>
<tr>
<td>Imbalance that favored treatment arm</td>
<td>12</td>
</tr>
<tr>
<td>Statistical adjustment of imbalances</td>
<td>10</td>
</tr>
<tr>
<td>Futility of agent by pooled analysisb</td>
<td>10</td>
</tr>
<tr>
<td>Subgroup found to be positive without providing baseline information</td>
<td>8</td>
</tr>
<tr>
<td>Noisy endpoint (e.g., ordinal analysis)</td>
<td>4</td>
</tr>
<tr>
<td>Unrepresentative control arm by pooled analysisb</td>
<td>3</td>
</tr>
</tbody>
</table>

Out of the 16 agents that were preceded by early-phase trials, all but tirilizad had at least one identifiable factor in an early phase that we suggest may lead to the erroneous interpretation of efficacy or safety. Their frequency is shown here; many trials had more than one factor identified. Table e-2 provides details for each individual early- and late-phase trial.

---

a Many trials had more than one factor identified.

b Many trials did not have sufficient information to compare to the pooled outcome models.
3. Given the small sample sizes in early-phase trials, dichotomous outcomes appear to be more reliable, while ordinal outcomes are more appropriate for secondary hypothesis generation and not to reduce projected sample size as has been suggested.

4. The use of historical controls, matched through individual datasets or compared to surrogate outcome models, could provide a rapid way to identify early signals. A better sense of natural history will support whether the specific trial’s control arm is representative of a larger population and would likely be replicated when recruitment occurs in more diverse locations.

5. The use of valid post hoc rebalancing methods should be considered if the underlying assumptions are verified and the characteristics of the excluded subjects are explicitly provided.

6. Because noise, disease variability, and baseline imbalances and the inability to account for them contributes to false early-phase results and the failure to successfully identify effective agents, these factors necessitate a higher threshold for strength of effect to reasonably expect replication in a multisite study.

Although this review focused on the authors’ area of interest, ischemic stroke, the issues apply more broadly. Indeed, the Alzheimer disease and amyotrophic lateral sclerosis clinical research communities have recognized the importance of better understanding of natural history to properly design trials of therapeutic intervention.\(^5\quad 4\)\(^5\) While there is room for improvement in all aspects of translational medicine to identify and validate clinically effective agents, our suggestions are directed at the important question of progression beyond early- to late-phase trials and the investments that it requires.

**AUTHOR CONTRIBUTIONS**

Dr. Kent: study concept, design, and critical revision of the manuscript.

Dr. Shah: acquisition of data.

Dr. Mandava: study concept, design, and statistical analysis.

**STUDY FUNDING**

Drs. Kent and Mandava were funded by a pilot grant from Baylor College of Medicine, Institute of Clinical and Translational Research (T.A.K.), and the Department of Veterans Affairs VISN 16 pilot grant program (P.M.). Computation resources of the Computational and Integrative Biomedical Research Center at Baylor College of Medicine were utilized for this study (P.M.).

**DISCLOSURE**

T. Kent is a copyright holder of pPREDICTS. T. Kent has no commercial interest in its use.

S. Shah reports no disclosures relevant to the manuscript.

P. Mandava is a copyright holder of pPREDICTS. P. Mandava has no commercial interest in its use. Go to Neurology.org for full disclosures.

**REFERENCES**


