A Method to Determine Stroke Trial Success Using Multidimensional Pooled Control Functions

Pitchaiah Mandava, MD, PhD, MSEE; Thomas A. Kent, MD

Background and Purpose—Many early phase trials in stroke have not been subsequently confirmed. Randomization balance in baseline factors that influence outcome are difficult to achieve and may be partly responsible for misleading early results. We hypothesized that comparison with an outcome function derived from a large number of pooled control arms would mitigate these randomization problems and provide a reliable predictor for decision-making before proceeding to later phase trials. We developed such a model and added a novel feature of generation of multidimensional 95% prediction surfaces by which individual studies could be compared. We performed a proof-of-principle study with published clinical trials, determining whether our method correctly identified known outcomes.

Methods—The control arms from all randomized, controlled trials for acute stroke with ≥10 subjects, including baseline National Institute of Health Stroke Scale, age, and 3-month outcomes published between 1994 and May 2008, were identified. A Matlab program (PPREDICTS) was written to generate outcome functions based on these parameters. Published treatment trials were compared with these 95% intervals to determine whether it successfully identified positive and negative trials.

Results—Models of mortality and functional outcome were successfully generated (mortality: \( R^2 = 0.69 \); functional outcome, modified Rankin Scale 0 to 2: \( R^2 = 0.81 \); both \( P < 0.0001 \)). The National Institute of Neurological Diseases and Stroke intravenous recombinant tissue plasminogen activator trial and 3 studies yet to be subjected to Phase III study had modified Rankin Scale 0 to 2 outcomes above the 95% prediction interval. Sixteen treatment arm outcomes fell within prediction surface bounds. This group included 2 major trials, Stroke-Acute Ischemic NXY Treatment and Abciximab Emergent Stroke Treatment Trial, that initially appeared promising but went on to negative Phase III results.

Conclusion—This proof-of-principle analysis confirmed all positive and negative clinical stroke trial results and identified some promising therapies. The use of a pooled standard treatment group function combined with statistical bounds may improve selection of early studies for further study. This method may be applicable to any condition in which baseline factors influence outcome and at any stage of the development process.

Key Words: acute stroke ■ cerebrovascular accident ■ clinical trials ■ drug trials ■ outcomes

A acute ischemic stroke remains a leading cause of death and disability. Therapeutic options to treat an acute stroke are limited and over 100 randomized clinical trials have taken place with only one US Food and Drug Administration-approved acute treatment resulting in intravenous recombinant tissue plasminogen activator that occurred over a decade ago. Several early clinical trials such as Abciximab Emergent Stroke Treatment Trial (AbESTT)\(^5\) and Stroke-Acute Ischemic NXY Treatment (SAINT) \(^\text{I}^\text{I}\) were considered promising only to be proven later to be no better than placebo (AbESTT-II,\(^\text{II}^\text{G}\) SAINT-II\(^\text{G}\)). This early apparent success and the failure of many trials may be due to a number of reasons.\(^2,7,8\) Among the reasons not widely considered include an imbalance in randomization of baseline factors that influence outcome\(^6\) or an unrepresentative sample in earlier trials.

We hypothesized that comparison to a pooled standard treatment function could minimize randomization errors that frequently occur in small series and provide a broad stroke population with which to compare outcomes. This method could also mitigate the need to compare individual noncontrolled trials with prior published case series despite dissimilar baseline characteristics.\(^10\)

We previously used a pooled control sample in assessing benefit from the aggregate of intra-arterial therapy results by comparison with a 3-dimensional model of mortality, including age and baseline National Institutes of Health Stroke Scale (NIHSS) as predictors developed by Uchino et al.\(^10,11\) Here, we extend this prior approach by surveying a much larger number of studies. This new method differs in an important way by the addition of the novel feature of 3-dimensional 95% prediction intervals, thus providing a...
means to assess the likelihood a given study is different from the pooled sample. We test the feasibility of this analytic concept and the usefulness of comparing an individual study’s outcomes against the model predictions at comparable baseline variables. We analyzed how studies fared compared with our functions including those with large follow-up trials as well as early results yet to be confirmed.

Methods

The overall design of the model had 3 phases: (1) literature search and database creation; (2) model development; and (3) testing individual studies against the model.

Literature Search and Database Creation

Medline databases were searched for the words “acute ischemic (or ischaemic) stroke” and “randomized controlled trial” in combination with names of drug or intervention. Studies lay within, below, or above the surfaces. For this analysis, we selected median NIHSS and mean age and mortality and/or modified Rankin Score (mRS) outcomes at 3 months and beyond were selected for further scrutiny. Corresponding authors were contacted for clarification if complete information was not provided. Characteristics of each study were entered into a data structure in Matlab R2008a. This database is extendable to include characteristics currently not analyzed and more restrictive criteria could be used if desired to address specific stroke populations.

Model Development

To stabilize variance of data presented as proportions and to account for variability due to different number of subjects, Freeman-Tukey modification of the arc-sine square root transformation was first applied to the outcome data. The arc-sine square root transformation permits the generation of Gaussian function from proportions (eg, the proportion that achieved mRS 0 to 2). Minimization and Outlier Elimination

Nonlinear optimization functions (nlinit and lscurvefit) supplied by Matlab were used for minimization of the function. Both functions use a n-dimensional Newton method called the “interior trust region approach” of minimization. As per recommendations, outliers were eliminated if the Studentized residual was >2. The minimization/outlier elimination routine was automatically iterated until no further outliers remained. The fitting and minimization routines were independently verified using SigmaPlot 9.0 to establish further confidence in custom Matlab functions.

Multidimensional Prediction Intervals

Efforts in the previous 2 steps produce a set of parameters associated with minimized function and generate an r² and a probability value. Output from the previous 2 steps was used in a Matlab function to generate prediction intervals for the function and prediction intervals for an observation based on principles outlined in the Engineering Statistics Handbook. Prediction intervals take into account noise in a future observation as well as noise in generation of the minimized function and the variability of the data used to model the function. This concept was expanded from calculating prediction intervals at a single point to generating surfaces in 3 dimensions to represent surfaces that bound the model function on either side. The final program (Pooled Placebo REsponse DICtates Treatment Success [PPREDICTS]) calculates expected outcomes and ±95% prediction surfaces. For this analysis, we selected median NIHSS and mean age for the baseline variables. The percent that died and achieved an mRS 0 to 2 were the outcome measures.

Testing the Model

Outcomes of the treatment (and selected placebo) arms of all published randomized, controlled trials in which baseline NIHSS, age, and 3-month mortality and/or mRS 0 to 2 were compared against the model function and the ±95% surfaces generated by the second and third previously described steps and determined which studies lay within, below, or above the surfaces.

Results

Generation of Models

Twenty-eight studies representing 7136 subjects satisfied inclusion criteria and their control arms were used for mortality model development (Table). In generation of the control mortality arm function, 4 of 28 studies, representing 140 (2%) of 7136 patients, had Studentized residuals greater than the cutoff value and thus were eliminated in generating the minimized function. The function generated from the fit of the Freeman-Tukey transformed data (R²=0.69; P<0.0001) is plotted as a surface along with the ±95% prediction interval surfaces (supplemental Figure IA; available online at http://stroke.ahajournals.org). The mean NIHSS and age of the final group of studies was 12.2 and 67.5, respectively.

Performance of the Randomized, Controlled Trial Treatment Arms Against the Predictive Functions

Mortality

Four treatment arm mortalities were higher than the +95% prediction interval (ATLANTIS [letter L], GAIN [letter N], Prolyse in Acute Cerebral Thromboembolism [PROACT-II; letter Z], and PROACT-III [letter b]), indicating greater than expected mortality (Figure 1A). Twenty-three treatment arm mortalities were within the prediction interval bounds indicating that mortality was no different than predicted natural history. The Middle cerebral artery Embolism Local fibrinolytic intervention Trial (MELT; letter U) had mortality below the −95% prediction interval, indicating lower than expected mortality.

Functional Outcome

The outcomes in 16 studies were within the prediction bounds indicating no different from controls. Four studies, Neuro-Thera Effectiveness and Safety Trial-1 (NEST-1; H), MELT (U), National Institute of Neurological Diseases and Stroke (NINDS; V), and CLOTBUST (X) had outcomes better than +95% prediction interval suggesting that their functional outcome was better than predicted from the pooled control treatment function (Figure 1B).

Trials Above +95% Prediction Interval: NINDS, NEST-1, CLOTBUST, and MELT

The percent of patients that achieved an mRS 0 to 2 in the NINDS treatment arm when plotted onto the prediction surfaces is above the +95% prediction interval (Figure 2A), consistent with the conclusion that intravenous recombinant tissue plasminogen activator within the 3-hour window is beneficial. When separately plotting NINDS Part 1 and Part
2, both are above the +95% prediction surface (supplemental Figure IIA). Had these been done sequentially, our method would have suggested that Part 2 should proceed. The NINDS trial generated some controversy because randomization was thought to have favored the treatment arm over the placebo arm.42 However, the percent of subjects in the NINDS placebo arm is near the prediction surface (Figure 2A), indicating that even in the presence of imbalances in strata of NIHSS, the overall result is representative of a broad population of patients with stroke.

Three trials yet to be subjected to Phase III results demonstrated outcomes above the +95% prediction interval of subjects that achieved an mRS 0 to 2, suggesting the probability that they would have improved outcomes compared with a large control group. The NEST-1 treatment arm used a low-energy laser to treat patients within 24 hours of an acute ischemic stroke. There were several baseline imbalances that mostly favored the placebo arm (ie, lower NIHSS and age). Despite these imbalances, when our model is applied to the data, the outcomes of the NEST-1 treatment arm is above the +95% prediction interval (supplemental Figure IIB) for a comparable control population. Note that the placebo arm outcome is below the prediction surface, suggesting it may not be representative of a broader population; hence, when replicated in a larger population, the strength of effect may be less in a follow-up larger study should the control arm be more representative.

CLOTBUST37 used transcranial ultrasound directed at a thrombus occluding a middle cerebral artery during administration of intravenous recombinant tissue plasminogen activator in a 3-hour treatment window. CLOTBUST treatment arm functional outcome is above the +95% prediction interval (supplemental Figure IIC) indicating it likely represents an effective treatment. Note that its control group, intravenous recombinant tissue plasminogen activator alone, is close but not above the interval. This different result from the clearly positive NINDS trial reflects a different population with a higher baseline NIHSS selected in CLOTBUST for large vessel occlusion. Interestingly, when the NINDS result was analyzed based on stratified NIHSS groups,43 it did
not demonstrate a substantial treatment effect in the NIHSS range of 15 to 20, the group represented here. This result may indicate that adjunctive ultrasound is particularly beneficial in more severe stroke when a less robust treatment effect from intravenous recombinant tissue plasminogen activator may be expected. CLOTBUST was also positive using mRS 0 to 1 as the outcome measure (as used in the original trial)\(^3\) (data not shown).

MELT\(^\text{44}\) used intra-arterial urokinase within 6 hours of onset compared with a control population who received an intra-arterial catheter and treatment with heparin. Our result indicates that this treatment was effective using mRS 0 to 2 as the outcome compared with this broader population (supplemental Figure IID). Its control group functional outcome is on the prediction surface, indicating it is representative of a broader population. The original publication indicated that treatment was effective with mRS 0 to 1 as the outcome, which was also confirmed in our mRS 0 to 1 model (data not shown). It is interesting that MELT restricted its maximum dose of urokinase to 600 000 U, a dose associated with better outcome in a prior study from the same group\(^4\) and virtually identical with that identified independently as optimal in our prior systematic review of the best performing intra-arterial therapies in which we compared outcomes with an earlier pooled sample.\(^1\)

**Early and Subsequent Phase III Results: SAINT and AbESTT**

SAINT-I,\(^4\) a large, early Phase III trial of NXY-059 in the 0- to 6-hour window, was thought to be positive in terms
of a “shift” in mRS but negative in terms of those that achieved a mRS 0 to 2. In our mRS 0 to 2 model (Figure 2B), SAINT-I and its control group outcomes were nearly identical to the control arm function. SAINT-II, a larger follow-up to SAINT-I, had outcomes that fell slightly below the predictive function surface, whereas its control group fell on the surface suggesting that this trial, like its precursor, was representative of the population at large. Neither result provides any signal for therapeutic benefit with this end point. AbESTT, the precursor trial to AbESTT-II and AbESTT-2C (companion cohort of AbESTT-II), and its control group were nearly identical with the control function and would not have appeared promising (supplemental Figure III). The subsequent larger AbESTT-II and AbESTT-2C and their control group functional outcomes were also close to the model surface confirming this lack of benefit.

**PROACT**

PROACT41 and PROACT-II, in which prourokinase administered through an arterial catheter was compared with intra-arterial heparin within a 6-hour treatment window in middle cerebral artery occlusion. The Phase III PROACT-II trial demonstrated a 15% improvement over control therapy that was positive by a stratified outcome method, but not by direct comparison of both groups (by our calculation using published data: \( P = 0.07 \); Fisher test of proportions). Imbalances that generally favored the placebo arm relative to the treatment arm were also noted in PROACT II. In our functional outcome models, PROACT-II was nearly positive in terms of mRS 0 to 2 (supplemental Figure IVA) and PROACT was positive with respect to mRS 0 to 1 (data not shown). However, both arms showed higher than expected mortality (supplemental Figure IVB) and the control arm had a substantially lower percent that achieved a good outcome.

**Figure 2.** The percent of subjects that achieved mRS 0 to 2 for the NINDS and SAINT-I and -II trials plotted onto the predictive function and ±95% prediction interval surfaces. A, NINDS intravenous recombinant tissue plasminogen activator treatment (blue dot) and placebo arm (red dot) outcomes. The percent that achieved an mRS 0 to 2 in the treatment arm exceeded the ±95% confidence interval. B, Treatment and control arm outcomes of SAINT-I (blue dot and blue circle) and SAINT-II (red dot and red circle). Both treatment and placebo arms clustered around the control outcome function, indicating little chance for treatment trial success for this end point.
by guest on July 13, 2017 http://stroke.ahajournals.org/ Downloaded from

ship among baseline NIHSS, age, and functional outcome

We were able to generate an excellent model for the relation-

mortality and lower good outcomes than a broader popula-

tissue plasminogen activator within 3 hours of stroke onset.2

Additionally, we identified 3 studies that performed better

apparent benefit of addition of prourokinase. Although it is

infusing heparin and prohibition of early administration of

PROACT, it is possible that use of an intra-arterial catheter

population or to control arms representing best medicine

may arise when the treatment is generalized to a larger

in PROACT, insights were gained into potential issues that

it is likely not possible to randomize for all variables with a

age) contributed considerable predictive value for outcomes,

although the factors considered here (baseline NIHSS and

could be a source for pooling this information. Note that

large public databases with detailed patient characteristics

accommodate this uncertainty. However, the availability of

outcomes and likelihood of subsequent success needs to

provide normative functions and it is likely that prediction

intervals will be broader. In that case, confidence in the

outcomes and likelihood of subsequent success needs to

accommodate this uncertainty. However, the availability of

large public databases with detailed patient characteristics

could be a source for pooling this information. Note that

although the factors considered here (baseline NIHSS and

gender, arterial territory, or treatment approaches provided there are suffi-

cient series published to permit generation of models. Al-

though it will not be possible to visualize beyond 3 dimen-

sions, a user is primarily interested in whether or not an

outcome of a trial at a particular point in n-dimensional

independent variable space is between the +95% and −95%

prediction values and therefore visualization is not critical. It

is our intention to add subsequent studies to the model on an

ongoing basis and test these additional factors and outcomes.

This way any potential drift in outcomes of the control arms

could be readily accounted for.

We envision 2 other potentials application of our method.

The first is to compare an individual study, with its own

confidence interval, on an ongoing basis with our functions as

subjects are accumulated. This would be particularly useful to

assess for early signals of increased mortality or benefit when

randomization imbalances are most likely to occur. Another

application is to estimate effect size. By comparison to a large

and diverse “control” group, we can assess whether a particu-

lar control group was indeed representative of what is to be

expected in a larger trial. In the control arm in NEST-1,23

43.9% achieved an mRS 0 to 2, whereas our method would

have predicted 48.4% ± 10%. This example indicates that our

method would suggest a more conservative estimate of effect

size in the planning process for follow-up confirmations. We

anticipate investigators will also use this feature to test early

case series in the absence of a control arm to estimate effect

size for subsequent randomized trials. In theory, our method

could be extended to any condition in which baseline factors

influence outcome in which imbalances in randomization

may lead to erroneous conclusions.

Disclosures

P.M. and T.A.K. hold the copyright for PPREDICTS and a patent

application has been submitted on their behalf for the method

described here. PPREDICTS, Inc has been formed by the Baylor

College of Medicine to commercialize these results.
References


A Method to Determine Stroke Trial Success Using Multidimensional Pooled Control Functions
Pitchaiah Mandava and Thomas A. Kent

Stroke. 2009;40:1803-1810; originally published online March 12, 2009; doi: 10.1161/STROKEAHA.108.532820

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/40/5/1803

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/