Efficacy Outcome Selection in the Therapeutic Hypothermia After Pediatric Cardiac Arrest Trials

Richard Holubkov, PhD[1]; Amy E. Clark, MS[1]; Frank W. Moler, MD, MS[2]; Beth S. Slomine, PhD[3,4]; James R. Christensen, MD[5,6,7]; Faye S. Silverstein, MD[2]; Kathleen L. Meert, MD[8]; Murray M. Pollack MD[9,10]; J. Michael Dean, MD, MBA[1]

Objectives: The Therapeutic Hypothermia After Pediatric Cardiac Arrest trials will determine whether therapeutic hypothermia improves survival with good neurobehavioral outcome, as assessed by the Vineland Adaptive Behavior Scales Second Edition, in children resuscitated after cardiac arrest in the in-hospital and out-of-hospital settings. We describe the innovative efficacy outcome selection process during Therapeutic Hypothermia After Pediatric Cardiac Arrest protocol development.

Design/Setting: Consensus assessment of potential outcomes and evaluation timepoints.

Interventions: None.

Measurements and Main Results: We evaluated practical and technical advantages of several follow-up timepoints and continuous/categorical outcome variants. Simulations estimated power assuming varying hypothermia benefit on mortality and on neurobehavioral function among survivors. Twelve months after arrest was selected as the optimal assessment timepoint for pragmatic and clinical reasons. Change

[1]Department of Pediatrics, University of Utah, Salt Lake City, UT.

[2]Department of Pediatrics, University of Michigan, Ann Arbor, MI.

[3]Department of Neuropsychology, Kennedy Krieger Institute, Baltimore, MD.

[4]Department of Psychiatry, Johns Hopkins University School of Medicine, Baltimore, MD.

[5]Department of Pediatric Rehabilitation Medicine, Kennedy Krieger Institute, Baltimore, MD.

[6]Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, Baltimore, MD.

[7]Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD.

[8]Department of Pediatrics, Wayne State University, Detroit, MI.

[9]Division of Critical Care Medicine, Children's National Medical Center, Washington, DC.

[10]Department of Pediatrics, George Washington University School of Health Sciences, Washington, DC.

ClinicalTrials.gov identifiers: THAPCA-OH (NCT00878644), THAPCA-IH (NCT00880087).

Supported, in part, by the Pediatric Emergency Care Applied Research Network (PECARN) under cooperative agreements U03MC00001, U03MC00003, U03MC00006, U03MC00007, and U03MC00008 from the Emergency Medical Services for Children program of the Maternal and Child Health Bureau of the Health Resources and Services Administration, and from the National Institute of Child Health and Human Development Collaborative Pediatric Critical Care Research Network (CPCCRN) under cooperative agreements U10HD500009, U10HD050096, U10HD049981, U10HD049945, U10HD049983, U10HD050012, and U01HD049934.

Dr. Holubkov served as board member for Pfizer and the American Burn Association (Data and Safety Monitoring Board [DSMB] memberships), consulted for St. Jude Medical and the Physicians Committee for Responsible Medicine (Biostatistical consultancies) and received support for article research from National Institutes of Health (NIH). Dr. Holubkov and his institution received grant support from National Heart, Lung, and Blood Institute (NHLBI; chief biostatistician for Therapeutic Hypothermia After Pediatric Cardiac Arrest [THAPCA]). His institution received support for travel from NHLBI (THAPCA planning meeting). Ms. Clark received support for article research from NIH. Her institution received grant support from the NIH. Dr. Moler received support for article research from NIH. His institution received grant support, support for travel, and support for participation in review activities (R21 HD044955 and R34 HD 050531 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development [NICHD], and by U01 HL094339 [Dr. Dean] and U01 HL094345 [Dr. Moler] from the NHLBI). Dr. Slomine received support for writing/reviewing the article from NHLBI (U01HL094345/co-investigator), received support from NHLBI (grant pays for administrative support and overhead), served as board member for the American Board of Clinical Neuropsychology (Travel expenses as oral examiner), consulted for the University of Michigan (Executive Committee for Planning Grant) and University of California, Davis (DSMB Member), is employed by Kennedy Krieger Institute, provided expert testimony for private practice, lectured for St. Joseph's Hospital (Presentation at Grand Rounds), and received support for article research from NIH. Dr. Slomine and her institution received support for travel from NHLBI (U01HL094345/co-investigator). Her institution received grant support from NHLBI (U01HL094345/co-investigator). Dr. Christensen is employed by Kennedy Krieger Institute and received support for article research from NIH. His institution received grant support, support for travel, and support for writing/reviewing the article from NHLBI (U01HL094345/co-investigator) and received support from NHLBI (grant pays for administrative support and overhead). Dr. Silverstein received support for travel from the March of Dimes (Scientific advisory board) and received support for article research from NIH. Her institution received grant support from NHLBI (funding for role as co-investigator on grant UO1 HL094345) and from NICHD (effort funded on an unrelated project HD073692) and received support for travel from NHLBI (investigator meeting HL094345). Dr. Meert received support for article research from NIH. Her institution received grant support from NIH. Dr. Pollack received support for article research from NIH. His institution received grant support. Dr. Dean's institution received grant support from NHLBI, NICHD, and NIH.

DOI: 10.1097/PCC.0000000000000272

Address requests for reprints to: Richard Holubkov, PhD, Intermountain Injury Control Research Center, 295 Chipeta Way, Suite 2E600, Salt Lake City, UT 84158. E-mail: rich.holubkov@hsc.utah.edu

in Vineland Adaptive Behavior Scales Second Edition from prearrest level, measured as quasicontinuous with death and vegetative status being worst-possible levels, yielded optimal statistical power. However, clinicians preferred simpler multicategorical or binary outcomes because of easier interpretability and favored outcomes based solely on postarrest status because of concerns about accurate parental assessment of prearrest status and differing clinical impact of a given Vineland Adaptive Behavior Scales Second Edition change depending on prearrest status. Simulations found only modest power loss from categorizing or dichotomizing quasicontinuous outcomes because of high expected mortality. The primary outcome selected was survival with 12-month Vineland Adaptive Behavior Scales Second Edition no less than two SD below a reference population mean (70 points), necessarily evaluated only among children with prearrest Vineland Adaptive Behavior Scales Second Edition greater than or equal to 70. Two secondary efficacy outcomes, 12-month survival and quasicontinuous Vineland Adaptive Behavior Scales Second Edition change from prearrest level, will be evaluated among all randomized children, including those with compromised function prearrest.

Conclusions: Extensive discussion of optimal efficacy assessment timing, and of the advantages versus drawbacks of incorporating prearrest status and using quasicontinuous versus simpler outcomes, was highly beneficial to the final Therapeutic Hypothermia After Pediatric Cardiac Arrest design. A relatively simple, binary primary outcome evaluated at 12 months was selected, with two secondary outcomes that address the potential disadvantages of primary outcome. (*Pediatr Crit Care Med* 2014; XX:00–00)

Key Words: cardiac arrest; clinical trials; hypothermia; randomized; simulations

Cardiopulmonary arrest is a catastrophic event associated with high mortality rates and with poor quality of life among many survivors caused by neurological injury. Several randomized trials have demonstrated long-term benefit of therapeutic hypothermia (cooling to core temperatures of 32–34°C for 12–72 hr) on survival and neurological outcomes. These trials were performed in adults resuscitated after sustaining cardiac arrest out of the hospital (1, 2) and in neonates less than 6 hours old presenting with hypoxic–ischemic encephalopathy (3, 4). Findings from these trials cannot be extrapolated to the large population of infants and older children experiencing cardiac arrest either out of the hospital (in settings such as near drowning) or in the hospital (often in settings of preexisting major illness). Additionally, there is concern about possible higher short-term mortality rates after therapeutic hypothermia in children because of a strong trend reported in a pediatric traumatic brain injury trial (5). Because of the lack of a well-powered trial assessing benefit of hypothermia in children resuscitated after cardiac arrest, our research group initiated the Therapeutic Hypothermia After Pediatric Cardiac Arrest (THAPCA) trials. These trials evaluate safety and efficacy of therapeutic hypothermia when compared with therapeutic normothermia (actively maintaining body temperature at 36–37.5°C to prevent fever) in two separate populations of pediatric patients. Because of differing etiologies, resuscitation quality, and causes of acute mortality between children sustaining cardiac arrest in the out-of-hospital versus in-hospital setting (6), as well as generally more rapid treatment initiation when arrest occurs in hospital, separate THAPCA trials will be performed in these two populations. A description of the rationale, study design, and protocol for the THAPCA trials has been published (7).

We describe here the clinical, logistic, and technical aspects of this process. Our practical experiences may inform the design of future critical care studies assessing outcomes combining survival and functional status among survivors.

METHODS

Consensus Process

The expert consensus process of selecting appropriate primary and secondary efficacy endpoints in the THAPCA trials involved 1 year of extensive discussion among acute care clinicians, neurobehavioral outcome specialists, and biostatisticians. At a "kick-off" organizational planning meeting in August 2006 attended by approximately 20 individuals, study outcomes including timeframe for follow-up were discussed along with other protocol aspects, but consensus regarding outcomes was not achieved. After various smaller group protocol development discussions, and expert input from the Collaborative Pediatric Critical Care Research Network and Pediatric Emergency Care Applied Research Network, the instrument and timing for assessing the primary outcome were finalized by expert consensus at a protocol development meeting in November 2006. Subsequent technical study outcome finalization, which included statistical simulations and other technical discussions described below, was facilitated by regular telephone conferences attended by the authors of this report. These conferences occurred from January to July 2007, at which time consensus was achieved regarding study endpoints.

Trial Design

As previously described (7), THAPCA consists of two parallel prospective multicenter randomized trials. Institutional Review Boards at all THAPCA centers approved the protocol and informed consent documents. Parental permission is provided for each subject.

RESULTS

Components of the Primary Efficacy Outcome

The benefit of therapeutic hypothermia, if one exists, may be on survival, on neurobehavioral status among survivors, or on both of these. A scenario where hypothermia is beneficial for one of these outcomes and detrimental for the other cannot be ruled out. Therefore, it was necessary that the primary THAPCA outcome measure incorporate both survival and neurobehavioral status and be robust to different magnitudes of treatment effect on each of these.

Neurobehavioral Assessment Measure

In THAPCA, children range from 2 days to 17 years old. All are comatose at time of randomization. Some surviving children

may be vegetative or severely disabled. Detailed neurobehavioral assessment must be performed using information provided by a parent or caregiver. Assessment of the child's function before the cardiac arrest is important because some children (particularly those who had cardiac arrest while hospitalized) will have had substantial preexisting neurobehavioral deficits. Prearrest function assessment must be obtained retrospectively from a parent, at a stressful time shortly after their child's cardiac arrest. Although masking parents to assigned treatment is not possible in THAPCA, it is highly desirable to obtain this assessment before parental knowledge of their child's initial response to assigned treatment.

The Vineland Adaptive Behavior Scales-II (VABS-II) (8)) was selected as the primary instrument for assessing neurobehavioral status. Unlike two other caregiver report measures of adaptive behavior considered, the Scales of Independent Behavior-Revised (9) and the Adaptive Behavior Assessment System-Second Edition (10), the VABS-II has only one version of the test for the THAPCA age range of 0 to 18 years, whereas the other two tests have different versions for children of varying ages. Therefore, the VABS-II allows more uniform comparison throughout the entire THAPCA age range. When compared with the other two measures considered, the VABS-II also has more items that capture behaviors of very young or low-functioning children, which is particularly important because there is the potential for many of the older children enrolled in THAPCA to be low functioning.

The VABS-II is appropriate for measuring neurobehavioral outcome from birth to adulthood, in children ranging from very low functioning (vegetative/minimally conscious) to fully functional and independent. It includes four domains (communication, daily living, socialization, and motor skills), each broken down into subdomains. Items within a subdomain are sequenced developmentally starting with skills typically observed at the youngest age. Its psychometric properties are strong because the VABS-II has been standardized on a large normative sample representative of the United States population. The VABS-II includes a parent-caregiver rating form, which is a rating scale format, and a survey interview form, which is designed as a semistructured interview. Importantly, there were no significant score differences in caregiver responses between these two form types in the standardization sample (8). The VABS-II survey interview is also suitable for centralized remote administration. Telephone administration has been validated versus in-person administration (11), interrater reliability is high (8), and a Spanish-language version exists for the survey edition.

Timing of Primary Outcome Assessment

There is evidence that neuropsychological function improves from the acute postcardiac arrest period to 6-month follow-up in adults (12); similar pediatric data do not exist. Yet, THAPCA investigators believed that it was important to measure the primary outcome at a delayed timepoint to allow for neurological recovery, and that 12 months was the earliest evaluation timepoint that would be considered a long-term behavioral outcome after cardiac arrest. Although later intervals, such as 18 months, were considered, 12-month evaluation would allow more patient enrollment within the study timeframe with lower loss to follow-up. In addition, pediatric follow-up data (13) showed significant improvement during the first year after traumatic brain injury, with subsequent plateauing of function. Consequently, the THAPCA investigators' pragmatic consensus decision was to select 1 year after cardiac arrest as the timepoint when neurological recovery would be relatively complete; most subjects would be medically stable, and high rates of subject enrollment, retention, and follow-up could be facilitated.

Outcome Assessment Logistics

To measure prearrest neurobehavioral function, the parent-caregiver rating form of the VABS-II is completed by caregivers of THAPCA participants shortly after randomization. At 3 months and at 1 year after randomization, the VABS-II survey edition is administered to parents by a small number of experienced telephone interviewers at a central facility (Kennedy Krieger Institute). Reliability between the parent-caregiver rating form and survey edition is extremely high (8). Interviewers are masked to treatment assignment and not otherwise in contact with patients' families. Given difficulties in transporting patients with complex medical conditions, it was anticipated that telephone-based interviews would yield higher follow-up rates than in-person visits. Having a small number of experienced interviewers performing telephone-based assessment centrally is also cost-effective and may limit between-interviewer variability.

The VABS-II assesses whether the child can perform a list of various tasks across domains. The number of each type of task that can be performed is standardized to the child's age using a reference normal population with mean score of 100 and SD of 15. As an artifact of this standardization to a mainly normal-functioning cohort, the lowest possible standardized VABS-II scores for very low-functioning children differ slightly according to age.

A standardized semiquantitative neurological examination, together with detailed neuropsychological testing, will be performed at THAPCA clinical sites among surviving children whose parents allow participation in these complementary assessments. These data, although very informative, are considered tertiary; the VABS-II will be used for main treatment effect assessment.

Specific Primary Outcome Selection

After consensus was achieved with respect to assessment instrument and timepoint, clinical, practical, and biostatistical issues were evaluated to determine the specific primary outcome for the THAPCA trials. From a clinical perspective, interpretability, reproducibility, and ability to generalize the outcome measure were paramount considerations. From a biostatistical perspective, issues including potential bias, missing data, and statistical power of the final comparison were considered. Two major issues influenced selection of the primary outcome: the impact of prearrest neurobehavioral status and attainment of optimal granularity (level of detail). Primary outcomes of six candidates are summarized in **Table 1** along with their strengths and limitations.

TABLE 1. Candidate Therapeutic Hypothermia After Pediatric Cardiac Arrest Efficacy Outcomes

Outcome	Strengths	Weaknesses
Outcomes assessing change from prearrest to 12 mo		
1. Quasicontinuous change score (death assigned lowest value, lowest possible VABS-II at one year next lowest value)	Highest statistical power/ granularity Adjusts for prearrest functional status	Prearrest VABS-II possibly missing/inaccurate Inappropriate to analyze as completely continuous Results of statistical analysis difficult to interpret clinically, as magnitude of change Magnitude and clinical significance of potential change vary according to baseline VABS-II
2. Multicategorical, five levels: 1) death; 2) lowest possible VABS; 3) worsening > 30 points; 4) worsening 16–30 points; 5) worsening ≤ 15 points	Improved power vs dichotomous outcome Clinically meaningful categories Adjusts for prearrest functional status	Prearrest VABS-II possibly missing/inaccurate Multiple cutpoints arguably subjective Some categories not achievable for children with low prearrest VABS-II Lowest possible VABS-II varies by age
3. Dichotomous (alive with worsening ≤ 30 points)	Relatively interpretable and clinically meaningful "single" outcome Adjusts for prearrest functional status	Prearrest VABS-II possibly missing/inaccurate Cutpoint arguably subjective Less statistical power because of limited granularity Children with baseline VABS-II < 30 points above minimum must be excluded
Outcomes assessing 12-mo status only		
1. Quasicontinuous status (death assigned lowest value, lowest possible VABS-II at 1 year next lowest value)	High statistical power and granularity Prearrest VABS-II not required	Power loss with no baseline adjustment Inappropriate to analyze as completely continuous Results of statistical analysis difficult to interpret clinically, as magnitude of effect
2. Multicategorical, four levels: 1) death; 2) VABS-II < 45 (includes minimally conscious/vegetative); 3) VABS-II between 45 and 69; 4) VABS-II ≥ 70	Improved power vs dichotomous outcome Uses clinically meaningful categories Prearrest VABS-II not required	Power loss with no baseline adjustment Multiple cutpoints arguably subjective
3. Dichotomous (alive with VABS-II ≥ 70)	Most interpretable and clinically meaningful "single" outcome Prearrest VABS-II not required for calculation	Power loss with no baseline adjustment Less statistical power because of dichotomization Cutpoint arguably subjective Children with baseline VABS-II < 70 must be excluded

VABS-II = Vineland Adaptive Behavior Scales-II

Change From Prearrest Status Versus 1-Year Status Alone

Because prearrest functional status is expected to be heterogeneous in the THAPCA populations, outcomes based on changes from prearrest level could more accurately capture treatment effect for each case and thus improve relative statistical power. Change-based outcomes would facilitate inclusion of children with poor prearrest neurobehavioral status, who

comprise a non-negligible proportion of eligible patients (particularly in the in-hospital setting) and who could not improve to a good level regardless of treatment efficacy. Excluding such children from the trial is ethically unacceptable.

However, "change from prearrest status" outcomes require accurate assessment of prearrest neurobehavioral status. The necessarily retrospective parental assessment of the child's prearrest status, performed in extremely stressful circumstances within 24 hours of cardiac arrest, is subject to inaccuracies and will not be available for some children. At the time of this assessment, parents are aware of the assigned treatment; nonetheless, parental recall or reporting biases should be equally distributed between treatment arms.

Another argument against "change from prearrest" outcomes is that a difference of a given magnitude in VABS-II scores is more disabling at lower levels. For example, a 20-point decrease from 80 (low average) prearrest to 60 (low) at 1 year will have greater adverse impact on functioning than the same 20-point decrease from 110 (average range) to 90 (still within average range). Maximum potential decline from prearrest level is also lower for children with compromised prearrest function. The child's ultimate functional status and capabilities after intervention may also be considered more important to parents and clinicians than magnitude of decrease from prearrest status.

Level of Detail

Using a quasicontinuous or multicategorical outcome would be expected to achieve higher statistical power than a binary outcome. Power gain, however, is limited by the expected high proportion of deaths in THAPCA subjects. For defining quasicontinuous outcomes, death is the worst status, and the lowest possible age-specific VABS-II score is next worst (incorporating vegetative or minimally conscious children). Children alive at 1 year without disorders of consciousness will be assessable using a continuously distributed measure, either 1-year VABS-II score or change in VABS-II score from baseline.

A practical weakness of quasicontinuous outcomes is quantifying overall treatment effect, over and above the statistical comparison between treatment arms. For example, a rank-based quasicontinuous outcome comparison might find a marginally significant overall treatment difference, with modest between-arm differences in both mortality and in VABS-II among survivors. Clinicians examining only the *p* value and summary data might be unsure of the magnitude, precision, and "location" of the treatment effect and thus be unconvinced of its practical importance. This limitation was one motivation for consideration of a multicategorical or binary primary outcome. **Table 1** includes two versions of multicategorical 1-year outcome for which consensus was achieved, one incorporating baseline status and another using only 1-year status. The 15-point and 30-point VABS-II increments were selected because calibration of VABS-II to a normal population incorporates 15-point SD. Disadvantages of categorical outcomes include compromised statistical power when compared with quasicontinuous measures and arbitrary determination of VABS-II category cutpoints. In addition, children with poor prearrest VABS-II scores are unable to achieve outcome categories corresponding to either favorable 1-year VABS-II levels or to substantial worsening of VABS-II from baseline, compromising interpretability of treatment effect for the entire population. Finally, the age-varying threshold for lowest possible VABS-II score could compromise interpretability of categories across the age spectrum.

The simplest outcomes considered were binary classifications of "survival with acceptable functional status at 1 year" and "survival at 1 year without substantial worsening from prearrest neurobehavioral status." There was substantial investigator consensus to define acceptable 1-year functional status as VABS-II score greater than or equal to 70. This cutpoint, two SDs below the reference population mean of 100, is considered a low level of functioning. For dichotomized change from prearrest status, a drop in VABS-II score more than 30 points from prearrest level, representing a change of two SDs in the reference population, was proposed. Combining death and poor/worsened functional status into a single category was considered acceptable by clinicians. These binary endpoints, particularly dichotomized 1-year status alone, were viewed as clinically interpretable, pragmatic, and sufficiently objective. Acknowledged limitations included possible loss of statistical power when compared with continuous and multicategorical measures and need to exclude cases with poor prearrest neurobehavioral function (e.g., VABS-II < 70) from the primary efficacy analysis.

Sample Size and Power Estimation: Technique and Assumptions

To estimate sample sizes required for acceptable statistical power, simulations were performed under various assumed treatment effects of hypothermia on survival and on neurobehavioral outcome among survivors. These simulations involved generating prearrest VABS-II scores for a cohort of children, simulating categorical 1-year status of mortality, vegetative/minimally conscious state, or survival without disorder of consciousness for each child, and further simulating a treatment effect on VABS-II for realizations where the child survived at 1 year.

A key assumption was the distribution of prearrest VABS-II scores. Pediatric Overall Performance Category (POPC) and Pediatric Cerebral Performance Category (PCPC) data were reviewed from a retrospective cohort study of children resuscitated after cardiac arrest that had been performed at 15 hospitals expected to participate in THAPCA (7). It was estimated that about 65% of in-hospital cases and 85% of out-of-hospital cases would come from a typically developing reference population (with normally distributed prearrest VABS-II scores with means of 100 and SD of 15). Remaining cases were simulated as arising from a generally impaired population (normally distributed VABS-II scores with mean of 70, and a wider range with SD of 20). For each simulation, every case was randomly selected as coming from either the normal or impaired population using a Bernoulli distribution. Any

TABLE 2. Outcomes and Event/Mortality Rates in Therapeutic Hypothermia Trials Used for Therapeutic Hypothermia After Pediatric Cardiac Arrest Parameter Estimation

Trial	Population	Outcome	Event Rate: Hypothermia Arm (Specific Treatment) (%)	Event Rate: Comparative Arm (Specific Treatment) (%)	Mortality: Hypothermia Arm (%)	Mortality: Comparative Arm (%)
Bernard et al (1)	77 adults treated within 2 hr of out-of-hospital ventricular fibrillation	Survival at hospital discharge, discharged home or to rehabilitation facility	49 (hypothermia, 12 hr)	26 ("normothermia")	51	68
Hypothermia After Cardiac Arrest (2)	275 adults treated 5–15 min after cardiac arrest	6-mo survival with favorable neurologic outcome	55 (hypothermia, 24 hr)	39 (conventional care)	41	55
CoolCap (3)	234 term infants with encephalopathy, treated within 6 hr	18-mo survival without severe disability	45 (hypothermia, 72 hr)	34 (conventional care)	33	37
National Institute of Child Health and Human Development Neonatal Network (4)	208 term infants with encephalopathy, treated within 6 hr	18- to 22-mo survival without severe disability	54 (hypothermia, 72 hr)	38 (conventional care)	24	37

generated prearrest scores below 20 (below achievable VABS-II values) were removed.

Changes from baseline status were then simulated. For each case, cutpoints applied to a uniformly distributed random variable determined death, vegetative/minimally conscious status, or survival without consciousness disorder at 1 year, per specified arm-specific probabilities. For patients surviving without consciousness disorder, "change from baseline VABS-II" was generated from a normal distribution, with SD of 15 points and mean determined by the hypothesized treatment effect. Any realizations with resulting 1-year VABS-II score of 20 or below were categorized as vegetative/minimally conscious. Finally, distributions of resulting quasicontinuous and ordered categorical outcomes were compared between treatment arms by an exact rank-based Wilcoxon test (14), whereas binary outcome rates were compared via standard chi-square test.

To increase potential power of between-arm comparisons, we also considered analyzing quasicontinuous outcomes as mixed distributions, partly categorical (dead and vegetative status) and partly continuous (1-year VABS-II or change in score), and simultaneously comparing the two components using likelihood-based approaches (15). However, a perceived disadvantage of this approach was "omnidirectionality," wherein a treatment that (for example) not only increases mortality but also improves function among survivors would have both the categorical and the continuous distribution components substantially different from the other treatment (resulting in a highly significant p value), despite no overall patient benefit when survival and function are considered together. An approach such as the Wilcoxon test that inherently and simultaneously ranks and compares all possible outcomes including mortality was judged to be more appropriate when comparing quasicontinuous outcomes between arms.

Estimation of treatment effects and survival rates was challenging because limited data were available from two out-of-hospital trials performed in adults (1, 2) and two trials in neonates (3, 4) (**Table 2**), and these populations differed substantially from THAPCA with respect to age and disease characteristics. For the simulations, the possible beneficial absolute effect of hypothermia on survival was estimated at 15% in the out-of-hospital setting and 10% in the in-hospital setting (where rapid intervention and immediate access to maximal care might limit hypothermia benefit). Possible beneficial hypothermia effect on neurobehavioral function in survivors was estimated to range from 5 to 15 points (i.e., up to 1 SD in a normal population distribution). Initial mortality estimates, based on acute mortality observed in the retrospective cohort study, were 50% for the in-hospital normothermic arm and 60% for the out-of-hospital normothermia arm.

Sample Size Simulation Results

Generation of simulated cohorts using the R package (16) was relatively simple computationally. Runs of 10,000 simulations were performed using a range of sample sizes, in increments of five subjects per study arm. **Table 3** shows minimum sample sizes required to achieve 80% and 90% power under various assumptions, for simulated in-hospital and out-of-hospital settings. Across simulations, although the sample size penalty for not incorporating baseline status into a particular outcome type ranged from nonexistent to nearly 80% in the in-hospital setting, this penalty was generally modest (usually under one third) in the out-of-hospital setting (where a stronger treatment effect on survival was postulated). The sample size penalty for using a

TABLE 3. Simulation Results: Sample Sizes Required Under Various Scenarios

Assumed Benefit of Hypothermia		Change Score: Quasicontinuous		Change Score: Multicategorical (Five Levels)		Change Score: Dichotomous (Dead/Δ VABS-II ≥ 30)		One-Yr Status: Quasicontinuous		One-Yr Status: Multicategorical (Four Levels)		One-Yr Status: Dichotomous (Dead/ VABS-II < 70)	
Survival	VABS-II	80% Power	90% Power	80% Power	90% Power	80% Power	90% Power	80% Power	90% Power	80% Power	90% Power	80% Power	90% Power
A. In-hospital THAPCA trial													
10%	5	380	530	430	570	500	650	530	700	590	780	700	940
10%	10	220	290	280	370	360	450	350	470	430	590	400	530
10%	15	140	190	210	280	290	370	250	310	350	460	270	350
B. Out-of-hospital THAPCA trial													
15%	5	220	300	230	310	250	320	250	330	270	360	320	420
15%	10	160	220	180	240	180	230	210	270	230	310	220	290
15%	15	130	170	150	210	150	190	170	220	210	280	170	220

THAPCA = Therapeutic Hypothermia After Pediatric Cardiac Arrest, VABS-II = Vineland Adaptive Behavior Scales-II.

Baseline VABS-II Distribution Assumptions: (A) 65% normal (mean = 100; SD = 15), 35% impaired (mean = 70; SD = 20) and (B) 85% normal (mean = 100; SD = 15), 15% impaired (mean = 70; SD = 20).

Normothermia arm assumptions: (A) 50% survival, decrease in VABS-II among survivors of −15 ± 15 points, 0.5% alive and comatose and (B) 40% survival, decrease in VABS-II among survivors of −20 ± 15 points, 5% alive and comatose.

Hypothermia arm assumptions: (A) SD of decrease in VABS-II of 15 points, 0% comatose and (B) SD of decrease in VABS-II of 15 points, 2.5% comatose.

categorical versus a quasicontinuous outcome was often appreciable in the in-hospital setting, mainly for outcomes accounting for baseline status, whereas the penalty for a less granular outcome was smaller in the out-of-hospital setting.

We identified simulation scenarios where a multicategorical outcome yielded inferior power to a binary outcome, particularly when a strong hypothermia effect was postulated on function among survivors. Because this observation was not immediately intuitive, we found it very instructive to examine actual proportions of patients with outcomes in each category observed in each simulation scenario (**Table 4**). In some scenarios assuming a strong hypothermia benefit on VABS-II among survivors, proportions in the "second best" category for each multicategorical outcome were higher in the normothermia than in the hypothermia arm, compromising power of a between-arm comparison of ordered multiple categories.

Required sample sizes were within the range of estimated numbers of eligible patients available for enrollment in the study time frame (700–900 across the two trials combined). Assuming at least a moderate effect of hypothermia on VABS-II scores in survivors, available patient numbers sufficed even with less granular outcomes. Overall, the THAPCA investigators believed that despite larger sample size requirements for binary outcomes, their simplicity and interpretability outweighed loss of statistical power relative to outcomes incorporating a higher level of detail. The binary outcome of survival with good neurobehavioral function was considered to be most relevant to parents and caregivers. Therefore, the primary THAPCA endpoint selected was survival with good neurobehavioral function (VABS-II ≥ 70) at 12 months after cardiac arrest. This outcome is meaningfully evaluable only among children with prearrest VABS-II greater than or equal to 70. Any children whose prearrest VABS-II is not assessable will be included (i.e., assumed to have sufficiently good prearrest VABS-II) if prearrest POPC and PCPC assessments both indicate at most mild disability.

Final Sample Size Calculations

For the primary binary outcome selected, investigators hypothesized that absolute hypothermia benefit would be higher (20%) in the out-of-hospital setting than in the in-hospital (15%). These estimates corresponded reasonably well with treatment benefit actually realized under the complex assumptions of the simulations used for power estimation (**Table 4**, "Dead or VABS-II < 70" column). Final sample size calculations were performed using standard methodology for a binary outcome, assuming the above magnitudes of treatment effect. A spectrum of possible outcome rates for the normothermia arm was estimated from the retrospective cohort study (7), which assessed general neurologic function using the PCPC (1, good; 2, mild disability; 3, moderate disability; 4, severe disability; 5, coma or vegetative state; 6, death). Children in the Severe Disability or Coma categories would have VABS-II scores below 70, and neurobehavioral expert investigators estimated that about half in the Moderate Disability category would have VABS-II below 70 (17). Resulting ranges of estimates for 12-month survival with VABS-II greater than or equal to 70 were 15% to 35% in the out-of-hospital normothermia arm and 35% to 55% in the in-hospital normothermia arm.

The final sample size requirements (**Table 5**) were based on a two-sided chi-square test comparing proportions with α=0.05 and incorporate a 2% inflation to account for interim Data Safety Monitoring Board efficacy monitoring using conservative O'Brien-Fleming boundaries (18, 19). On the basis of these calculations, final target sample sizes were set at 504 evaluable patients for the in-hospital trial (providing 90% power

TABLE 4. Proportions Observed for Candidate Outcomes Under Various Simulation Scenarios (Based on 10,000 Simulations, Each With 500 Patients Per Arm)

Arm/ Scenario	Binary: Dead or ΔVABS-II ≥ 30 (%)	Outcome Incorporating Baseline VABS-II (%)					Binary: Dead or VABS-II < 70 (%)	Outcome Using 1-Yr VABS-II Only (%)			
		Dead	Lowest Possible VABS-II	VABS-II Worsened ≥ 30	VABS-II Worsened 15–30	VABS-II Worsened < 15		Dead	VABS-II < 45	VABS-II 45–70	VABS-II ≥ 70
In-hospital THAPCA trial											
Hypothermia arm: 10% survival benefit; 5-point VABS-II benefit among survivors	46.1	40.0	1.1	5.0	16.2	37.7	60.4 52.5[a]	40.0	6.3	14.1	39.6
Hypothermia arm: 10% survival benefit; 10-point VABS-II benefit among survivors	43.3	40.0	0.7	2.6	12.0	44.7	56.7 48.7[a]	40.0	4.7	12.0	43.3
Hypothermia arm: 10% survival benefit; 15-point VABS-II benefit among survivors	41.7	40.0	0.4	1.3	8.0	50.4	53.5 45.7[a]	40.0	3.4	10.1	46.5
Normothermia arm: (identical across scenarios)	58.9	50.0	1.8	7.1	16.4	24.7	70.7 64.5[a]	50.0	7.3	13.4	29.3
Out-of-hospital THAPCA trial											
Hypothermia arm: 15% survival benefit; 5-point VABS-II benefit among survivors	56.1	45.0	3.1	7.9	17.7	26.2	63.9 60.4[a]	45.0	6.5	12.4	36.1
Hypothermia arm: 15% survival benefit; 10-point VABS-II benefit among survivors	52.5	45.0	2.9	4.6	14.4	33.1	60.3 56.8[a]	45.0	5.4	10.0	39.7
Hypothermia arm: 15% survival benefit; 15-point VABS-II benefit among survivors	50.1	45.0	2.7	2.4	10.6	39.2	57.3 53.8[a]	45.0	4.5	7.8	42.7
Normothermia arm: (identical across scenarios)	74.0	60.0	5.6	8.4	13.1	12.9	78.6 76.5[a]	60.0	8.7	10.0	21.4

THAPCA = Therapeutic Hypothermia After Pediatric Cardiac Arrest, VABS-II = Vineland Adaptive Behavior Scales-II.
[a]Rate considering only patients with baseline VABS-II ≥ 70 in each simulation.

to detect a 15% treatment effect in all settings) and 250 for the out-of-hospital trial (providing at least 85% power to detect a 20% treatment effect in all settings, with higher power if favorable outcome rates are relatively low as expected).

Selection of Secondary Outcomes

Secondary efficacy outcome selection was based on two main considerations: inclusion of children with prearrest VABS-II scores below 70 who were excluded from the primary analysis and incorporation of outcomes that would more clearly delineate any treatment benefits on survival versus improved VABS-II performance. Thus, one secondary efficacy outcome will be survival at 1 year, to be compared between treatment arms as a proportion, and with survival curves presented as a supportive analysis. An additional secondary efficacy outcome selected was change from prearrest status, analyzed as

TABLE 5. Sample Sizes Required[a] for Therapeutic Hypothermia After Pediatric Cardiac Arrest Trials Using a Binary Primary Outcome of Survival With Vineland Adaptive Behavior Scales-II greater than or equal to 70

Assumed Rate of Survival With VABS-II ≥ 70 in Normothermia Arm (%)	In-Hospital Scenario: 15% Hypothermia Benefit 80% Power	In-Hospital Scenario: 15% Hypothermia Benefit 85% Power	In-Hospital Scenario: 15% Hypothermia Benefit 90% Power	Out-of-Hospital Scenario: 20% Hypothermia Benefit 80% Power	Out-of-Hospital Scenario: 20% Hypothermia Benefit 85% Power	Out-of-Hospital Scenario: 20% Hypothermia Benefit 90% Power
15	274	312	360	170	192	222
20	312	352	402	190	214	246
25	340	386	446	204	230	264
30	362	410	474	214	240	278
35	376	426	494	220	248	286
40	384	434	504	222	250	288
45	384	434	504	220	248	286
50	376	426	494	214	240	278
55	362	410	474	204	230	264

VABS-II = Vineland Adaptive Behavior Scales-II.

[a]Estimated sample sizes assume a two-sided χ^2 test with type I error of 5% and reflect inflation for conservative interim monitoring as described in text.

quasicontinuous in a rank-based fashion, with death and vegetative/minimally conscious status treated as the respective worst-possible and next-worst-possible values for this change regardless of prearrest VABS-II. This outcome was selected to elucidate the greatest possible detail regarding treatment effect of hypothermia on improved function among survivors, while maintaining integrity of the randomization by including nonsurviving children. To facilitate interpretation, the rank-based comparison of this outcome will be accompanied by a table of distributions of the multicategorical outcome incorporating change (**Table 2**) by study arm. Because the two secondary efficacy outcomes were judged to be of equal importance, both comparisons will be performed using a α level of 0.025, incorporating a Bonferroni–Holm stepdown procedure (20) to maximize power.

DISCUSSION AND SUMMARY

In the planning of the THAPCA trials, investigators first achieved consensus that the VABS-II was an appropriate instrument to assess outcome in the study population, and that 12 months after cardiac arrest was the optimal evaluation timepoint from both pragmatic and clinical perspectives. Next, to determine the specific primary outcome, a spectrum of candidate outcomes ranging from quasicontinuous to binary were considered. This more technical element of the outcome selection process included extensive discussion between clinicians and biostatisticians about assumptions and expectations regarding population parameters and treatment effects. After a range of reasonable assumptions was determined, simulation studies quantified loss of statistical power associated with using less granular measures and with not incorporating prearrest functional status into the endpoint. These simulations demonstrated that needed sample sizes were practically feasible even with outcomes using a lower level of detail. Once this feasibility was established, simplicity, availability, and direct interpretability of the study outcome became paramount. The THAPCA primary outcome, 12-month survival with VABS-II greater than or equal to 70, was ultimately selected based on these considerations. Secondary outcomes were then selected to complement limitations of the primary outcome regarding inclusion of all randomized patients and detailed treatment effect assessment.

Assumptions regarding prearrest VABS-II distributions in the THAPCA populations, and magnitudes of treatment effect on survival and neurobehavioral function, were imprecise. This limitation was recognized and was one reason that basic power calculations for a binary outcome, rather than results of the more complex simulation studies, were used for final power justification.

Although the primary outcome selected was relatively simple, confidence regarding its use was only established after extensive simulations quantified its relative performance, demonstrated its feasibility with available sample sizes, and showed that magnitude of treatment effect generated under relatively complex assumptions was in line with results observed in prior trials. Subsequent selection of appropriate secondary outcomes was relatively unproblematic because advantages, drawbacks, and performance characteristics of each candidate outcome had been comprehensively addressed during the discussions and simulations.

Overall, we believe that the iterative collaborative outcome determination process implemented in the THAPCA trials worked very well. We hope that our experiences provide

insights for others planning trials where outcome timing, granularity, interpretability, and other performance issues are being considered.

REFERENCES

1. Bernard SA, Gray TW, Buist MD, et al: Treatment of comatose survivors of out-of-hospital cardiac arrest with induced hypothermia. *N Engl J Med* 2002; 346:557–563
2. The Hypothermia After Cardiac Arrest Study Group: Mild therapeutic hypothermia to improve the neurological outcome after cardiac arrest. *N Engl J Med* 2002; 346:549–556
3. Gluckman PD, Wyatt JS, Azzopardi D, et al: Selective head cooling with mild systemic hypothermia after neonatal encephalopathy: Multicentre randomised trial. *Lancet* 2005; 365:663–670
4. Shankaran S, Laptook AR, Ehrenkranz RA, et al; National Institute of Child Health and Human Development Neonatal Research Network: Whole-body hypothermia for neonates with hypoxic-ischemic encephalopathy. *N Engl J Med* 2005; 353:1574–1584
5. Hutchison JS, Ward RE, Lacroix J, et al; Hypothermia Pediatric Head Injury Trial Investigators and the Canadian Critical Care Trials Group: Hypothermia therapy after traumatic brain injury in children. *N Engl J Med* 2008; 358:2447–2456
6. Moler FW, Silverstein FS, Meert KL, et al: Rationale, timeline, study design, and protocol overview of the therapeutic hypothermia after pediatric cardiac arrest trials. *Pediatr Crit Care Med* 2013; 14:e304–e315
7. Moler FW, Meert K, Donaldson AE, et al; Pediatric Emergency Care Applied Research Network: In-hospital versus out-of-hospital pediatric cardiac arrest: A multicenter cohort study. *Crit Care Med* 2009; 37:2259–2267
8. Sparrow S, Cicchetti D, Balla D: Vineland Adaptive Behavior Scales. Second Edition. Minneapolis, MN: Pearson Assessment, 2005
9. Bruininks RH, Woodcock RW, Weatherman RF, et al: Scales of Independent Behavior - Revised. Itasca, IL: The Riverside Publishing Company, 1996
10. Harrison PL, Oakland T: ABAS II. Adaptive Behavior Assessment System, Second Edition. San Antonio, TX: PsychCorp, 2003
11. Limperopoulos C, Majnemer A, Steinbach CL, et al: Equivalence reliability of the Vineland Adaptive Behavior Scale between in-person and telephone administration. *Phys Occup Ther Pediatr* 2006; 26:115–127
12. Sauvé MJ, Doolittle N, Walker JA, et al: Factors associated with cognitive recovery after cardiopulmonary resuscitation. *Am J Crit Care* 1996; 5:127–139
13. Jaffe KM, Polissar NL, Fay GC, et al: Recovery trends over three years following pediatric traumatic brain injury. *Arch Phys Med Rehabil* 1995; 76:17–26
14. Hothorn T, Hornik K: exactRankTests: Exact Distributions for Rank and Permutation Tests. R package version 0.8–25. Available at: http://CRAN.R-project.org/package=exactRankTests. Accessed July 11, 2014
15. Lachenbruch PA: Comparisons of two-part models with competitors. *Stat Med* 2001; 20:1215–1234
16. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at: http://www.R-project.org/. Accessed July 11, 2014
17. Fiser DH: Assessing the outcome of pediatric intensive care. *J Pediatr* 1992; 121:68–74
18. O'Brien PC, Fleming TR: A multiple testing procedure for clinical trials. *Biometrics* 1979; 35:549–556
19. Jennison C, Turnbull BW: Group Sequential Methods With Applications to Clinical Trials. Boca Raton, FL: Chapman and Hall, 2000
20. Holm S: A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; 6:65–70