CLINICAL INVESTIGATION

Statistical design and analysis issues for the ARDS Clinical Trials Network: the Coordinating Center perspective

Clin. Invest. (2012) 2(3), 275–289

We describe the statistical design principals for clinical trials conducted by the Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network that is funded by the National Heart Lung and Blood Institute. These include the definition of ARDS used by the network, the choice of measures of treatment efficacy, the choice of sample size, eligibility criteria, the choice of data to collect, stratification and randomization, the use of factorial trials and coenrollment of patients on more than one trial, early stopping for futility and efficacy, the choice of control group, the control of cointerventions, treatment administration and covariate adjustment.

David A Schoenfeld*, Douglas Hayden, Cathryn Oldmixon, Nancy Ringwood & B Taylor Thompson

Department of Biostatistics, 50 Staniford Street, Suite 560, Massachusetts General Hospital, Boston, MA 02114, USA
*Author for correspondence:
Tel.: +1 617 726 6111
E-mail: dschoenfeld@partners.org

Keywords: acute lung injury • coenrollment • cointerventions • early stopping rules • factorial clinical trial • hospital mortality • on target • usual care • ventilator-free days

The Acute Respiratory Distress Syndrome (ARDS) Clinical Network was established by the the National Heart Lung and Blood Institute in 1994 to develop and conduct clinical trials to prevent, treat and improve the outcome of patients with acute lung injury (ALI), ARDS and possibly other related critical illnesses. The network has refined its approach to clinical trials over the past 16 years. The purpose of this article is to describe the network's approach to the major issues in study design and analysis, which are particular to trials of acute life-threatening diseases, such as ARDS.

The National Heart, Lung, and Blood Institute created the network by requesting proposals for each site that would enter patients and for a separate 'Clinical Coordinating Center' that would organize the network, provide clinical coordination and handle statistical and data management issues. This article focuses on the central statistical decisions made by the network; many of which were discussed in the original Coordinating Center application. Where possible we use the data accumulated by the network to evaluate these decisions. Other aspects of the network's organization are described elsewhere [1]. Table 1 summarizes the network studies to date [2–10].

The definition of ARDS

The original Coordinating Center proposal noted that there was not a single definition of ARDS in the literature [11–13], although all of the definitions agreed that ARDS needed to be defined as lung injury that was severe, acute and had no other obvious etiology. One of the first decisions of the network was to extend its scope from ARDS to ALI, with ALI defined as the first occurrence of a $PaO_2/FiO_2 < 300$, bilateral infiltrates and the requirement for positive pressure ventilation via an endotracheal tube all present in a 24-h time period. We decided that patients with $200 \leq PaO_2/FiO_2 < 300$ would be recruited for our trial along with patients with $PaO_2/FiO_2 < 200$. If the P:F ratio increases out of range in the interval prior

FUTURE SCIENCE part of fsg

10.4155/CLI.12.14 © 2012 Future Science Ltd

ISSN 2041-6792

Table 1. Summary of Acute Respiratory Distress Syndrome Trials Network studies to date.

Study	Design	Samples size	Primary end point	Treatment	Ref.
Low tidal volume	Factorialized	861	180-day mortality	6 vs 12 mg/kg V_T	[2]
Ketoconazole	Factorialized	234	Mortality	Ketoconazole vs placebo	[2,3]
Lisofylline	Factorialized	235	Mortality	Lisofylline vs placebo	[4]
Steroids	Prospective, randomized, placebo-controlled	180	60-day mortality	Methylprednisolone vs placebo	[5]
High PEEP	Prospective, randomized	549	60-day mortality	Higher PEEP/lower FiO_2 vs lower PEEP/higher FiO_2	[6]
Fluids and catheter	Factorialized	1001	60-day mortality	PAC vs CVC and liberal vs conservative fluid management	[7,8]
Albuterol	Prospective, randomized, placebo-controlled	282	Ventilator-free days	Albuterol vs placebo	[9]
Fish oil/early vs late feeding	Factorialized	282 (Omega)/ 1000 (EDEN)	Ventilator-free days	Omega 3 FA vs comparator and early vs late enteral nutrition	[10]
Statins	Prospective, randomized, placebo-controlled	1000 (planned)	60-day mortality	Crestor® vs placebo	

CVC: Central venous catheter; FA: Fatty acid; PAC: Pulmonary artery catheter; PEEP: Positive end expiratory pressure; V_T: Tidal volume.

to initiation of study procedures, the patient is still eligible for enrollment based on the fact that the lung injury is still present.

We evaluated this decision in our first study on low-tidal volume ventilation [2], and showed that mortality for those in the 200–300 group was 20.5 compared with 32.8% for those individuals in the <200 group. Although the mortality in the 200–300 group was not as great as that of the <200 group, it was still considerable and given that the pathophysiology of these conditions were the same, we continued to use $200 \leq PaO_2/FiO_2 < 300$ as a criteria for inclusion.

The network wanted to exclude hypoxemia due to left atrial hypertension (LAH). At the time, the accepted method of excluding LAH was measurement of the pulmonary artery occlusion (wedge) pressure with a pulmonary artery catheter that would have to be in place at the time of screening. If a wedge pressure was used as a eligibility criteria it would be necessary to place a pulmonary artery catheter in patients without one after obtaining informed consent. The network decided to adopt the recommendation from the American European Consensus Congress [14] to allow a subjective decision on whether a patient had LAH if they did not have a pulmonary artery catheter. If a pulmonary artery catheter was in place, eligible study participants needed to have a wedge pressure no < 18 mmHg.

The decision not to require a pulmonary artery catheter probably helped accrual of patients as a subsequent observational study showed that a pulmonary artery catheter may be harmful [15], and our own study showed that it was not necessary [7] for safe fluid management. This study, termed FACTT, where we placed a pulmonary artery catheter in half the patients, allowed us to approximate how many patients would have been excluded had we required a wedge pressure threshold.

Of the 435 patients that were randomized to a pulmonary artery catheter in FACTT, 128 had a wedge pressure of greater than 18 mmHg. Thus we could expect to include approximately 30% of patients with a high wedge pressure. However, the mean cardiac index in those with a high wedge pressure was also elevated consistent with a so called ‘hyper-dynamic circulation.’ This physiologic state may be caused by aggressive fluid administration creating volume overload in patients with normal or increased heart function, the latter from the stress of critical illness or intravenous catecholamines. There were only ten patients (~2%) that had both reduced cardiac function (defined as a cardiac index less than 2.5 $l/min/m^2$ body surface area) and a wedge of greater than 18 mmHg indicating systolic heart failure. Based on these estimates, our American European Consensus Conference-based inclusion criteria exclude systolic heart failure but not an elevated wedge pressure from volume overload. It is unclear how many patients may have had an elevated wedge pressure as the sole cause of their pulmonary edema and were thus misclassified as having ALI. Absent a sensitive and specific biomarker for ALI, we cannot know this proportion. However, half of the patients with an elevated wedge had values of 19 or 20, and in the presence of pneumonia, sepsis, and trauma (present in 89% of patients in FACTT), it was the opinion of the investigators that the majority had volume overload and ALI together.

Measurement of treatment efficacy in ARDS treatment trials

■ Measurement of mortality

The obvious measure for determining efficacy in Phase III ARDS treatment trials is mortality. In clinical trials the mortality rate has been measured at 30, 60, 180 days or 1 year after randomization, each giving different estimates of mortality. In some papers, the times of death are compared. The original coordinating center proposal suggested using total 180-day mortality as the primary end point and this was used in the first trial. However, the network decided not to determine vital status of patients who were discharged home without mechanical ventilation. The rationale was that if they had subsequently died it was most likely from a cause other than ARDS. Furthermore, the long-term follow-up would be difficult for many of the centers that were tertiary care hospitals. Even though subjects were not contacted once they were home, the 180-day time horizon was a burden to study staff because this patient population was sometimes transferred to other hospitals or to skilled nursing facilities. Therefore, the data were not in the ARDS Network institution's medical record. In addition, such a long time period resulted in a substantial delay in acquiring completed data.

Our first study of lower tidal volume ventilation (ARMA) allowed us to evaluate the decision to use 180-day mortality. Only ten patients out of 898 died between 60 and 180 days. Thus changing the time horizon to 60 days would have changed the mortality rate by 1.1%, an error of 3%. If we had changed to 28-day mortality the error would have been 7%, which we considered more serious. In all but one subsequent study, we used 60- rather than 180-day mortality. The study that continued to use 180-day mortality was a study of steroids for patients who had been on a ventilator over 14 days and where prolonged ventilation was expected.

The decision to consider any patient who returned home off ventilation as living could be considered controversial, as we might be misclassifying some patients who died from ARDS. We have two sources of data to study the consequences of this decision; the first, a trial of lisofylline, conducted with an industrial sponsor, in which the US FDA required that patients be followed for treatment-related adverse events for 60 days after leaving the study hospital, whether or not they went home. Of the 153 patients who were discharged home, only one patient subsequently died before day 60. The second study involved a separately funded long-term follow-up study of ARDS survivors. This study followed a subgroup of our patients for up to a year and attempted to obtain mortality data from official sources if the patient could not be reached by telephone [16]. In this study, the 60-day mortality was 27%, which presumably included patients who went home and was comparable to hospital mortality reported in the primary study (fluid management: 26.95%; pulmonary artery catheter vs central venous catheter: 26.85%). A direct count of the number of patients whose death was determined after discharge home is not possible because patient level data on the long-term follow-up study is not available. The fact that mortality was the same in the follow-up study where patients who went home were followed, would lead one to believe that at 60 days we are not losing many deaths by failing to contact patients at home.

The data from the long-term follow-up study gives a somewhat different impression of the mortality of patients from 60 to 180 days than the data from our first study, with the mortality rate increasing from 25 to 29%. If we combine the two sources of data it would imply that an additional 3% (4–1%) of patients die after discharge home between 60 and 180 days. We do not currently have access to the data from the follow-up study so we cannot count this number directly.

The decision whether to follow patients at home after discharge brings about a scientific and a practical question. The scientific question is should we include all deaths in our calculation of mortality or should we attempt to exclude patients who died of other causes by excluding deaths that occurred after the patient's recovery from ARDS as measured by their ability to return home off ventilation. The argument for including these deaths is that it is never certain that a death is really unrelated to a subject's treatment during a trial. The argument for excluding them is that if they are truly unrelated, including them increases the required sample size of the study. The sample size increase that is required by increasing the mortality rate with unrelated deaths is relatively small. Suppose, for example that the death rate on placebo went from 25 to 30% because we included deaths that were not affected by the study treatments; the proportional increase in sample size to compensate for the extra 'noise' would then only be approximately 12%.

The decision to use 60-day mortality and not contact patients at home should be revisited periodically as both scientific and practical considerations may change. If it were easy to follow patients who went home, the certainty that we had accounted for all the deaths that might be affected by treatment would be worth the extra sample size and the coordinating center would have argued for including these deaths. Given our current difficulty in following these patients and the fact that we might lose quite a few to follow-up, we agreed to consider patients who return home off ventilation as alive. The advent of a national medical record may fundamentally change our cost–benefit analysis by making it easier to follow patients after they go home.

■ Statistical methods to compare mortality

The two most common methods used to compare mortality are the log-rank test [17], which utilizes the patients' time of death, and the test that compares the proportions of patients who died before the time horizon of 60 or 180 days. The log-rank test actually tests whether patients in one group die before patients in the other. Our original proposal suggested that the proportion dying before 180 days be used because the goals of ARDS treatment are to increase the number of patients who survive ARDS, rather than increasing the amount of time that it takes them to succumb to it [18]. A longer course in the ICU that ends in death is not a desirable outcome from our point of view. In addition, the actual survival time may be modified by the treating physician when he determines when care is withdrawn from a moribund patient.

We conducted a power analysis to see if we lost power by ignoring the time of death. Suppose that an effective treatment increases the proportion of patients, say by 15%, who will recover from ARDS but does not affect the time of death of patients who ultimately succumb. The power of the log-rank test when a proportion of patients survive and the time until death for the remaining patients is exponential was calculated [19] and compared with the power of a test comparing two proportions.

A trial using the log-rank test would require 4% more patients per arm than a test based on the proportion who survive 180 days, depending on the trial duration (power = 0.8). The log-rank test is not advantageous because the information that distinguishes the treatments, regardless of whether or not patients recover, is known with near certainty by 180 days. The information about when patients who die during the 180 days will be largely random variation. Other common tests that use the time of death would give similar results. Thus we use the Fisher exact test for the analysis of the principal study outcome. If there are important covariates we would use the stratified version of the Fisher exact test [20] or a logistic model. When the data are incomplete at an interim analyses, we use the Kaplan–Meier estimates at 60 days; alternatively, we use a test that notes when patients go home [21].

■ Ventilator-free days

At the initial ARDS Network meeting we considered another end point: ventilator-free days (VFD). This is defined as the number of days after the last day of mechanical ventilation to day 28. If a patient dies during the first 28 days, they have zero VFD. It is an end point that combines information about the duration of mechanical ventilation in survivors and mortality [22]. Patients who die do not get credit for unassisted breathing. Thus this end point makes more sense than total duration of ventilation or duration of ventilation in survivors, both of which ignore the mortality rate on each treatment.

The choice of how to compare VFD between the treatment groups is somewhat arbitrary. The paper that discusses the method [22] suggests using a nonparametric Wilcoxon rank sum test, while other ARDS trials have used a t-test. There is a false perception that data that are not normally distributed need to be analyzed using a nonparametric test. For large studies, the central limit theorem guarantees the validity of the t-test for most non-normal distributions [23]. An advantage of parametric analysis is that the test statistic are proportional to the estimated treatment effect so that one cannot have the situation where the estimated effect is very small while the test is significant, or the estimate is large and the test is not significant. Furthermore, there are easy techniques for correcting for covariates and conducting sequential analyses when parametric methods are used.

Whether the t-test is as powerful as the Wilcoxon test depends on the distribution of VFD and how it is affected by treatment. In the FACTT there was a significant difference in VFD between the fluid liberal and fluid conservative treatment arms [8]. In this study, the Wilcoxon test would have been more efficient requiring 26% fewer patients to show the same effect; in our KARMA study comparing ketoconazole and placebo it would have required 11% fewer patients. Thus, in terms of power, the Wilcoxon would have been superior.

The usual estimator to use with the Wilcoxon test is the Hodges–Lehman estimate, the median of all pair-wise differences between the treatments. In the KARMA study there was an average difference of two VFD. The Hodges–Lehman estimate of the difference was zero despite the fact that the difference was highly significant. The reason for this is that 30% of the observations are tied at zero, so the median difference occurs at zero. The difference is significant because 57% of the non-zero pairs are greater than zero.

The estimated difference in VFD is somewhat confusing, whether the Hodges–Lehmann estimate is used or the difference of the means. The problem is that the unit is days but the estimate combines mortality and the duration of ventilation in survivors. For instance, in the FACTT the difference in VFD was 2.43 days favoring conservative fluid management, the difference in the duration of ventilation among survivors was nearly the same, 2.56 days [8]. In our trial of lower tidal volume ventilation [2] the difference was 2 days but the difference in duration of ventilation among survivors was only 1.07 days.

In the fluid study, the difference in VFD and the difference in the duration of ventilation among survivors were nearly equal because there was not a large survival benefit. In the tidal volume study the difference in the duration of ventilation was much smaller than the difference in VFD because much of the difference in VFD was due to the mortality difference.

A drawback of using VFD as an efficacy measure is the potential danger that a study treatment will facilitate early weaning from mechanical ventilation but will prolong the patient's recovery of other functions or cause long-term harm other than mortality. In the FACTT, VFDs were significantly improved by conservative fluid management but the survival difference that also favored conservative fluid management was not large or significant. In terms of short-term measures of patient benefit, the data showed that the conservative fluid management strategy was not harmful and decreased the duration of ventilation. We had less information on long-term outcomes. Two long-term outcome studies were conducted on a relatively small fraction of the FACTT patients and will increase our understanding of the benefits and harms of this strategy.

The choice of what is the primary and what is the secondary end point is a help in decision making when the results of a trial are equivocal, but should not obscure the interpretation of strong clinical trial results. VFD were a secondary end point in the FACTT; the 3% improvement in mortality was not significant. Despite this we interpreted this trial as a positive study. The difference in VFD had a p-value of < 0.001 and all other indicators of improvement in duration of illness were also significant. We believe that there may also be a small mortality difference that would be difficult to detect.

Sample size for ARDS clinical trials

Determining the sample size of a clinical trial is the most imprecise activity that a statistician ever attempts. The formulae are simple and the calculation is easily performed using tools on the internet [101]. The problem, however, is that the determination of the size of the difference to be detected is so subjective that sample size calculations are often made backwards; the sample size is determined based on practical considerations and the detectable treatment difference is calculated from the sample size.

Conceptualizing differences in mortality rates involves a certain amount of difficulty. Suppose we hope to reduce a 40% control group mortality rate to a 30% rate. This can be seen as a 25 or 10% decrease depending on whether you wish to think of relative reductions or absolute ones. The English language is confusing on how to describe a difference in percentages. We refer to a reduction from 40 to 30% as a 10-point reduction because a 10% reduction could also mean a change from 40 to 36%. To detect a 10-point reduction, with 90% power, requires approximately 1000 patients. Suppose we subsequently learn that the control rate is larger or smaller than 40%. How should we modify the sample size? The answer depends on whether we considered the reduction from 40 to 30% as a 25% decrease in mortality or a 10-point reduction in mortality. Figure 1 shows the required sample size for either a 25% or a 10-point reduction when the 40% control mortality rate either decreases or increases. When the control group mortality rate decreases to 20% the number of patients needed to detect a 25% decrease in mortality (from 20 to 15%) doubles and the number of patients needed to detect a 10-point difference (from 20 to 10%) is cut by nearly half. Thus if we 'fix' the difference to be a 25% decrease, we need more patients as we improve the mortality of the control group; while if we 'fix' the difference at a 10-point difference, fewer patients are needed.

This is nearly what has happened during the 15 years

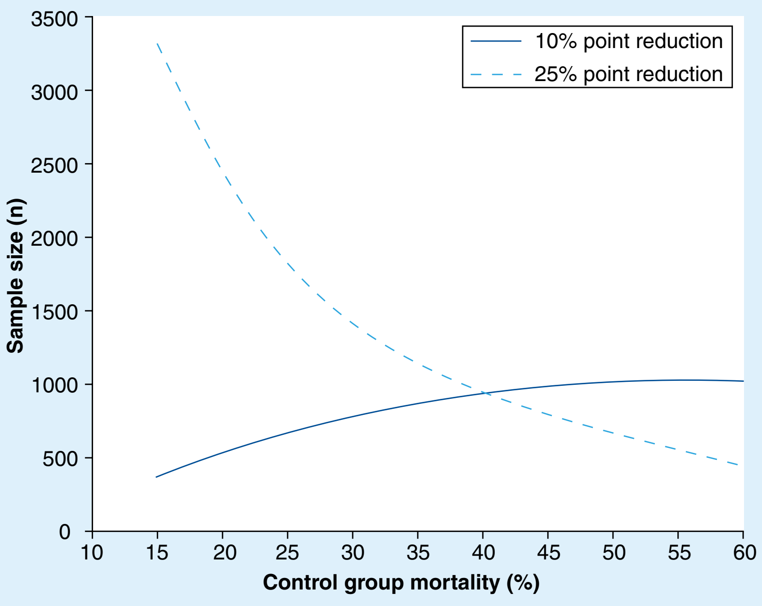

Figure 1. Affect of control group mortality on sample size.

of the ARDS Network. We started in 1996 with a 40% mortality rate for ARDS in our control group. Over the ensuing 15 years the mortality rate in the study population has dropped to nearly 20%. Based on this new mortality rate we should be using a sample size of 2000, or 500 depending on whether we conceptualized our goal as a 25% decrease or a 10-point difference. Instead, we have kept on using a sample size of 1000, which can detect approximately a 7–8-point decrease in mortality based on our current mortality rates. We hope this discussion clarifies some of the difficulties that sample size decisions entail. In practice, the decision on sample size consideration is a complex combination of opportunity cost versus chance of success. If we conduct larger trials that can detect very small differences we will do fewer trials and possibly detect fewer effective treatments. If we do smaller trials we may detect no differences at all. It is hard to say that a mortality difference of any magnitude is not clinically important. Furthermore, with a fixed number of centers, a larger trial would have to go on longer and after a point it is hard to maintain enthusiasm for a clinical study. As the network is currently organized, a clinical trial of 1000 patients is the largest trial that appears to be feasible in a 3- to 4-year time period. If we need to detect smaller differences more centers would be required.

A discussion of sample size would be incomplete without mention of our study on the use of steroids for patients with more than 5 days of mechanical ventilation who appeared not to be recovering [5]. When we initiated that study the thinking was that such patients would be common, but when we began to conduct the trial we found that the number of eligible patients was small. When a study has difficulty accruing patients the question being posed is of less clinical interest because it affects only a small population. This argues for stopping such a study due to futility or trying to detect a larger clinical difference that would have the same public health consequences as the smaller difference that was originally hypothesized. Although it is difficult to reconstruct all the decisions made at that time, the investigators, still blinded to the results of the study, chose to reduce the sample size and recalculate the power based on a larger hypothesized difference.

All of the ARDSNet studies are designed with futility and efficacy stopping rules so our sample size of 1000 is a maximum sample size. Most studies will stop earlier. This is discussed in a later section of this paper.

Efficacy measurements based on lung function

In our original coordinating center application we hypothesized that ARDS severity would be assessed several times during the course of treatment, but that these data should not be used as a primary measure of treatment efficacy because decreases in ARDS severity, as measured by physiologic parameters, would not predict improved long-term survival. Rather, it might be analyzed to give an insight into how an effective treatment works or why a treatment was ineffective. A simulation was conducted as to whether differences in lung injury scores might be used as an efficacy measure. The simulation showed that the sample size would not be decreased by this strategy.

These predictions were supported by the data that was collected, in that the effect of treatment on study measures of lung function has been inconsistent. For instance, the $Pa0_2/FiO_2$ ratio was lower in the superior 6 mg/kg group in our first study [2] and higher in the superior fluid conservative group as seen in our fluid study. The same is true of the lung injury score. In our ALVEOLI study (a negative study of higher vs lower positive end expiratory pressure [PEEP]) [6] the P/F ratio was much better in the higher PEEP group. This was an expected outcome because PEEP improves oxygenation by increasing the mean airway pressure, without necessarily improving the pathophysiology of ARDS.

Determination of eligibility criteria

The eligibility and exclusion criteria for an ARDS treatment trial should be chosen using the following general principles:

- Patients must have ARDS using a generally accepted and scientifically justified definition that can be implemented in a clinical trial;
- If testing a clinically available treatment, patients must not have a medical condition that absolutely requires the use of the treatment being tested;
- Patients must not have a contraindication to any of the treatments being tested;
- There must be a possibility that the patient could benefit from the treatment. For instance, patients with an extremely high risk of death from non-ARDS-related causes would be excluded as would those who have a very small risk of death or prolonged ventilation.

It is important to avoid making eligibility criteria so restrictive as to prevent rapid accrual or compromise the generalizability of the study. This is accomplished by being careful about how the last two principles are applied in practice.

In our first study over 90% of the following exclusions were due to the patient's medical condition: intracranial pressure (31%), chronic lung disease

(23%), terminal illness (17%), chronic liver disease (9%), bone/lung transplant (6%), neuromuscular disease (4%) or age <18 years (3%). Increased intracranial pressure was considered a contraindication for low tidal volume due to the possibility of the patient developing a high CO_2 level. The other exclusions, except for age, were conditions that might prolong mechanical ventilation or cause death that was unrelated to ARDS treatment. The exclusion of children was due to the fact that childhood ARDS has a much better prognosis than adult ARDS and is considered by some a different disease.

Of all the patients screened, 25% were excluded because consent could not be obtained within 36 h of disease onset. The exclusion of patients who have been on the ventilator for a good length of time has been our most problematic exclusion criteria due to the fact that it requires a rapid identification of patients and consent process. We feel that an ARDS treatment is most likely to be effective if it is applied early in the course of the illness. If many patients have had a long treatment before beginning the protocol therapy we would have less chance of showing an effect.

Determination of the data to be collected

One problem in clinical trials is the quantity of the data that must be collected, keyed and stored. Each additional datum adds to the cost and complexity of the clinical trial and may make it less likely that the necessary datum will be recorded accurately. Data collected in clinical trials can be divided into three categories; data that are absolutely necessary for the analysis of the clinical trial, data that might be useful for assessing the biological effects of the treatment or designing future trials, and data that are used for documentation and are usually not reported.

The data that must be collected in a clinical trial are data that indicate the efficacy of the treatment, data on important baseline characteristics and data on adverse events. Peto *et al.* urge that these be the only data collected in a clinical trial [24]. An additional use of data is to document the extent to which the protocol is followed and as an aid to improve protocol compliance. Our approach to protocol adherence is described in further detail later in this article.

For a Phase III trial of ARDS treatment the data collected would show the date the patient died or left the hospital alive, what illness led to the patient's ARDS, and measures of baseline covariates and adverse events. The baseline covariates could be the variables used to calculate the APACHE II score [25], which we collect rather than the score itself. Serious adverse events are reported directly on an adverse event form.

Data from clinical trials are often analyzed as an observational study; some examples are noted in the following publications [26–29]. The need to have data for secondary analyses often conflicts with the need to restrict the data in a clinical trial to avoid a burden on the study investigators, coordinators, data managers and site monitors. Where possible, secondary studies should be considered before the trial so that the necessary data elements can be collected. We have done this in our recent trials capturing alcohol use using the Alcohol Use Disorders Identification Test (AUDIT) scale [30]. In general, however, most of the data collected in clinical trials are not used for secondary studies. One problem in critical care trials is that the usual definition of a serious adverse event [102] leads to copious reports of events that are common in an ICU. In order to avoid having to constantly file adverse event reports we decided to capture organ failure systematically on our case report forms and then only require the emergent reporting of unexpected, serious, possibly study-related adverse events. This policy has been accepted by the FDA when we have conducted studies requiring FDA oversight. Based on this definition, the rate of reportable serious adverse events has been approximately one per 100 patients.

For Phase III trials questions of biologic effect are secondary. Measurements that fall into this category are those variables that determine the APACHE II–III score, other hematology and clinical chemistry measurements, urinalysis, chest film reports, ventilator parameters and frequency of infections. The question to ask when considering each item is what scientific hypotheses will be tested and how will they be analyzed. It is also important to remember that systematically collecting events that are either rare or ubiquitous usually does not provide useful information. We proposed reducing the expense of collecting these data by limiting the number of times these measurements are recorded.

Often clinical trials collect data for which the sole use is documenting end points or adverse events. The most extensive data in this category are data on drug dispensing, data from periodic physical examinations, data on concomitant medications and details on infections and other concurrent illnesses. The ARDS Network in general does not collect these data.

In our most recent trial involving statins we collect the data necessary to calculate the APACHE III score at baseline. We also use the AUDIT questionnaire to calculate alcohol dependence for a substudy and collect liver function tests and creatinine kinase levels for assessing statin liver toxicity. In addition, we collect vital signs, Glasgow coma information [31] and ventilation parameters at baseline at days 1–4, 7, 12, 21 and 28 while the patient is mechanically ventilated and we have a daily form to capture organ failures using the Brussels criteria [32]. Only a few specified concomitant medications are recorded without dosage details. Our safety reporting includes directly capturing both protocol- and nonprotocol-specified adverse events. Finally, there is a detailed study termination form that captures the primary and secondary end points of the study.

Stratification & randomization

Strata are groups of patients defined by patient characteristics that are thought to be prognostic. Stratification is the process of randomizing patients so that the number of patients in each treatment is near equal within each stratum. Stratification is often confused with using stratified statistical tests where prospectively identified patient characteristics are taken into account in the statistical analysis to correct imbalances and improve power. In large studies stratification is not necessary [24]. Most of the benefits that are attributed to stratification are actually benefits of stratified statistical tests. In our initial proposal we did not plan to use stratification because it unnecessarily complicated the randomization process.

Our randomization process naturally stratified patients by hospital because it used permuted blocks within each institution in order to facilitate drug trials where hospital pharmacies dispensed the medication. In addition, the ARDS Network's recent trials also stratified patients by whether they were in shock at study entry. We have not stratified our statistical analysis by hospital. To stratify by hospital, one would have to calculate the mortality difference within each hospital and then pool these differences across hospitals, rather than ignoring the hospital in our analysis as we have been doing. The problem can be thought of as follows: if the hospitals vary in their mortality rates then our estimated variance for the treatment difference is larger than it should be if we do not account for the hospital differences [30]. On the other hand, if we do account for these differences than in the beginning of the study we lose information based on how we account for these differences. We also increase the complexity of the analysis and make it nearly impossible to present the raw data that was the source of our summary statistics. The p-values for different choices of analysis method are a function of the measured difference divided by its precision and are a rough measure of the extent to which the precision is understated by an analysis that does not account for hospital. For instance, if the two methods of analysis lead to the same p-value then the precision was not understated in an analysis that ignored hospital. In our case, stratifying by hospital would introduce 50 strata. The benefits depend on how large the institutional effect is. For example, if the mortality rate from ARDS varies from hospital to hospital then stratifying by hospital will improve the power of a clinical trial. In relative terms the costs of stratification are reduced as the trial gets larger so they may be most problematic at the first interim analysis.

As an example of how this might work, we reanalyzed the KARMA trial with mortality as an end point; first stratifying the analysis by hospital and then doing the analysis without stratifying by hospital. The results were as follows: with the full sample size of 861 patients, the p-value was 0.0037 when stratifying by hospital, as compared with 0.0071 when not stratifying by hospital (22 hospitals in total). In terms of efficiency this would translate into a 22% increase in efficiency, that is, one could get the same power with 22% fewer patients. To show the effect of stratification with a small sample size, we also analyzed the mortality of the first 200 patients of KARMA; the p-values were 0.0068 and 0.0076, respectively, a 3% increase in efficiency. Thus, although it would not have made a difference in the KARMA study, there would have been more power in an analysis that stratified by hospital.

To show how stratifying by hospital might affect the analysis of VFD, we reanalyzed the FACTT (to compare the conservative and liberal strategies of fluid management) with VFD as the end point. With the entire sample of 1000 patients, the p-value when stratifying by hospital and not stratifying was the same, 0.0002 (41 hospitals in total). With the first 200 patients, the p-value was 0.23 when stratifying by hospital, in comparison to 0.05 when not stratifying by hospital. In this comparison it seems there is less advantage in analyzing the data stratified by hospital.

Factorial trials & coenrollment

In our initial proposal we argued that it would be advantageous to conduct trials that tested more than one hypothesis using a factorial design; a design where patients were randomized between four treatments. In such a scenario, patients received both treatment A and B, or patients received A but not B, B but not A, or neither A nor B. In such a design the effect of A is tested by stratifying the analysis on whether the patient received B or not B and similarly for testing the effect of B. We used this design for the network's first trial and it has since been implemented on nearly every trial the network has conducted. At the first meeting of the Network Investigators there was a controversy between those investigators who thought that our best chance to improve mortality was to find an effective drug and those investigators who thought that it would be best to use a lung protective ventilation strategy. The solution was to conduct a factorial trial. Patients were randomized into four groups: treatment with low tidal volume or treatment with a higher tidal volume and treatment with ketoconazole or placebo. Both therapies were tested in a single clinical trial that was the same size as a trial to test only one of the therapies.

Factorial trials require the assumption that there is no interaction between the two treatments. An example of an interaction would be a toxicity that was more likely if the patients received both treatments or a synergy or antagonism between the effects of the treatments. Factorial trials are particularly useful when the therapies act by different mechanisms and have different toxicities. Interactions are unlikely when it is not probable that both treatments will be effective. Factorial designs are also useful when supportive care is being studied because patients will have different treatments whether or not a factor is under study. Higher order factorial designs are possible with more than two treatments. Problems with these are that patients must be eligible for all the treatments, the treatments must not interact, and the trial must not be too hard to explain to the participants and institutional review boards.

The decision to have trials with more than two arms raises the issue of whether to adjust for the two treatment comparisons that will be made. We have not corrected for multiplicity in our analysis of factorial trials. The rationale is that if the two hypotheses were tested in separate trials no multiple comparison procedure would be used [33].

An alternative to doing a factorial trial is coenrollment, which has been used extensively in trials of cancer and has been described in the context of AIDS clinical trials [34]. Coenrolled trials allow the patient to be in two trials simultaneously provided they consent to both. For instance, we conducted a trial of albuterol versus placebo and simultaneously a factorial trial of a medical food versus a placebo and early versus delayed feeding. If a patient was eligible to be in both trials and would consent to both trials we obviously enrolled them in both. Otherwise, unlike a factorial trial, the patient could be in one or the other of the trials. Our case report forms were similar for the two trials and duplicate data did not have to be entered for coenrolled patients. Factorial trials are more cost effective than coenrolled trials but may be harder to enroll because patients must be eligible for both arms. Figure 2 shows our use of factorialization and coenrollment on network trials. Each of the trials is connected to the other trials for which it shared patients.

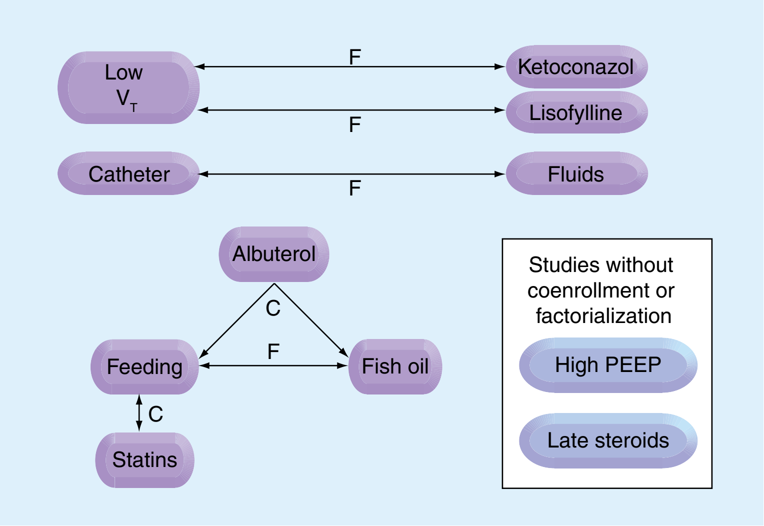

Figure 2. Coenrollment and factorial designs in Acute Respiratory Distress Syndrome Clinical Trials Network trials.
C: Coenrollment; F: Factorial; PEEP: Positive end expiratory pressure; V_T: Tidal volume.

Early stopping rules

Most clinical trials of fatal diseases conducted today use group sequential stopping rules. The primary justification is that it is not ethical to treat patients on a treatment that has been shown to be inferior or to enter them on a clinical trial that is futile. In addition, there are cost savings in stopping a trial early. The basic method [35] is that a trial has multiple analyses where the evolving data is evaluated by an independent Data and Safety Monitoring Board. The criteria used to decide whether to stop the trial at each analysis for efficacy are designed to give a 5% overall probability of a positive trial under the null hypothesis. Usually the criteria used in early analyses are chosen so that the trial has a very small chance of stopping early for efficacy under the null hypothesis [36] and the p-value for testing significance at the end of the trial is only slightly less than 0.05.

Usually these methods are conducted using the log-rank statistical test to compare treatment groups at each analysis. The log-rank test is not the most powerful test of whether the long-term survival rate is greater in one treatment than the other as was shown previously. Accordingly we base our early stopping rule on the normalized difference between the Kaplan–Meier survival curves at whatever time we are using to define mortality. The usual methods for finding group sequential boundaries were modified to use this statistic. After we know the survival status of every patient, this statistic becomes the difference in the proportions of patients that are alive at the time horizon chosen (now 60 days).

Table 2. Actual and maximum accrual for Acute Respiratory Distress Syndrome Clinical Trials Network.

Study	Number accrued (n)	Number planned (n)	Result
Low tidal volume	861	1000	+
Ketoconazole	234	1000	-
Lisofylline	235	800	-
High PEEP	549	750	-
Fluid management	1001	1000	+
CVP vs PAC	1001	1000	-
Albuterol	282	1000	-
Omega (omega-3)	282	1000	-
Eden	1000	1000	-

CVP: Central venous pressure; PAC: Pulmonary artery catheter; PEEP: Positive end expiratory pressure.

It became apparent to the network that stopping trials that were on a negative trajectory was also needed. In our first trial there was concern that there had not been a Phase II trial of ketoconozole of identical design to that we were proposing and yet we were planning a Phase III trial. Thus we adopted a Phase II/III roll-over design. We decided that after 200 patients were enrolled we would stop the trial if the results were such that we would not have chosen the drug for Phase III consideration had there been a 200 patient Phase II trial with those results. Since this stopping rule had not been used previously we developed the necessary statistical methods to do this [37]. Table 2 compares the actual accrual with the maximum accrual for the ARDS Network clinical trials. Most stopped early and our total accrual is approximately two-thirds of the maximum accrual.

The disadvantage of stopping for futility is that a trial that stops early will have a wide confidence interval around the estimate of the treatment effect. Thus the trial may provide little support for the statement that the new treatment is not effective. In some studies this is a minor problem while in other studies it is a major concern. Another disadvantage is that it is difficult to interpret the p-value and the treatment effect estimate when a trial stops. Four methods for correcting this p-value are given in Jennison and Turnball [38], which unfortunately may give different answers.

Most of the ARDS Network trials were designed to have three interim analyses (after each 250 patients) and a final analysis. The futility stopping rule can be characterized by how likely the trial will stop early if the null hypothesis is true. In order to more easily describe the rule we have used, we focus on the probability of stopping at either the first or the second look. Table 3 shows the cumulative probability of stopping for each of our trials. In the ketoconazole trial we had no need to demonstrate that ketoconazole was not effective. If the trial appeared to be futile it should be stopped. On the other hand, in the fluid and catheter trial, we were attempting to determine whether or not the choice of catheter was important and whether or not the fluid strategy was important so this trial was not designed to stop for futility at all. The statin and tidal volume studies were considered to be indeterminate in our need to accrue data to support the null hypothesis. For boundaries that have little chance of stopping early see the paper by DeMets and Ware [39], while in the Schoenfeld paper boundaries that tend to stop very early are described [37].

Coenrolled and factorial studies have been designed so that each factor could stop independently. For instance the randomization the ketoconazole or placebo stopped after approximately 200 patients while the lower versus higher tidal volume part of the trial continued. Stratified tests were used for dealing with the resulting mix of patients that one has after one or the other of the trials are stopped and for adjusting for coenrolment of patients.

Choice of a control group

Many of the ARDS Network trials have involved supportive care strategies such as tidal volume or management of fluids. The first step is to look at how these strategies are currently used. If there is wide variation in how a given treatment is used and if this variation is unexplained by patient factors, then the variation reflects clinician preferences or practice styles. In general, large unexplained variation indicates the lack of sufficient clinical evidence to guide a more uniform practice.

For the tidal volume trial, surveys of clinicians revealed wide variation in tidal volume preferences, subsequently confirmed when baseline (usual care) tidal volumes were recorded in our first trial [2,40]. Two Phase II trials revealed that very low tidal volume ventilation appeared to be beneficial and safe, yet was uncommon in usual care practice [41,42]. We chose to compare lower tidal volume (the experimental arm) to a fixed higher tidal volume that was both recommended at the time and represented what most clinicians preferred (the traditional approach). An alternative strategy suggested by some after the trial was to use usual care as the control, reasoning that clinician preferences may be better than either treatment group. However, studies that have usual care control arms have problems. A positive trial can be explained by the fact that some of the patients treated with usual care received poor care and a negative trial by the substantial overlap of the experimental arm and the control arm. Usual care is also vulnerable to secular trends and can change the nature of the trial over time with unpredictable results. Three-arm trials containing usual care and two fixed approaches have also been suggested, but these trials markedly increase the size of the trial. We simulated a number of three arm trials and found that the addition of a usual care arm increased the overall number of deaths in the trial before superior treatments could be identified under most assumptions [43].

Accordingly, our strategy has been to identify superior practices by comparing two potentially beneficial treatments that are both considered prudent approaches for the patient with ARDS. This determination is made by external and independent review by a Protocol Review Committee and the Data Safety Monitoring Board. When individual patient factors can be identified upon which to customize these approaches further, these features are built into our treatment arms.

Determination of the control is always controversial. Those that consider the control arm treatment to be extreme will assume that a positive trial is the result of a harmful control. A negative trial can be explained by a ‘U’ shaped dose response curve [44]. There was controversy about our interpretation of our trial of lower tidal volumes, but currently the use of lower tidal volumes has become commonplace [45].

Strategy for cointerventions

The treatment of an ARDS patient involves many factors including control of the ventilator, administration of fluids, feeding, sedation and paralysis, prevention and treatment of infections, and a myriad of other decisions that must be made by the medical staff of an ITU. The network had to decide how detailed the protocol should be in specifying how the patient should be treated. We call these other treatments cointerventions.

One argument for carefully controlling cointerventions is that the more variables you control the smaller the variance of the outcome measure. The proportional decrease in the variance is equal to the square of the correlation coefficient between the outcome and the covariate. For mortality, this effect is quite small, and is an average effect if you chose the cointervention at random. Suppose a cointervention changed the underlying mortality from 30 to 40%, the squared correlation coefficient between mortality and the choice of the cointervention is less than 1%. Therefore, controlling a cointervention that has a 10% effect on mortality has a negligible effect on the power.

Given these considerations, the network decided that cointerventions would be controlled for when there was evidence that the decision would improve patient outcomes. After we demonstrated that low tidal volume and conservative fluid management was beneficial, all subsequent studies controlled these factors. However, after our negative study on the use of corticosteroids, future use of corticosteroids was left to the investigator. The cointerventions of feeding and sedation are not specified in our protocols given the apparent lack of evidence to base our choice of treatment.

Quality control of treatment administration

In addition to more traditional methods of monitoring quality control of the study treatment (such as site visits), the network developed an innovative approach to quality control monitoring using on-target reports. We provided each site with a random time each day to collect clinical data relevant to the protocol. The advantage of using a random time was that the on-target percentage estimated the proportion of time that our patients were on target. Using these data, we developed an algorithm that determined whether an investigator was in compliance with protocol procedures and if the patient was on target related to the ARDS Network ventilator management protocol, weaning protocol, and more recently to the FACTT fluid management protocol. Currently, we generate monthly reports that compare each institution’s compliance and on-target rates with the on-target rates of the whole network over time. In addition, we send each institution a list of their off-target patients and the reasons why they were considered off target.

Clinical trials usually fall between two extremes in how the clinical protocol is enforced. On the one extreme are the studies that are designed to test a general strategy as it is applied in real life and on the other extreme are the studies designed to test a very specific intervention. Neither extreme is ever completely realized. Clinical trials are not real life nor is it possible to control everything. For this reason negative studies are often explained by insufficient differences between groups due to a lack of protocol compliance while positive findings are criticized

Table 3. Cumulative probability of stopping for futility at the first and second look under the null hypothesis.

Study	Stop at first look (%)	Stop by second look (%)
Low tidal volume	22	52
Ketoconazole	64	79
Fluids and catheter	0	0
Statins	6	25

because the treatments tested may not represent what is actually done in practice.

Controlling for prognostic factors

Our primary analyses do not control for prognostic factors. This is controversial as controlling for these factors may increase the power of the study [46,47]. The disadvantages of controlling for prognostic factors are that the factors that you control for must be recorded on every patient and mortality differences must be reported in terms of odds ratios rather than difference in percentages, which are easier to understand. We have used an analysis that controls for prognostic factors as a secondary analysis, which has been presented in most of our papers. The method we have used is described in detail below.

Logistic regression was used to derive a list of seven factors prognostic for mortality in the KARMA study. These factors, that were used as covariates in subsequent studies, are $AaDO_2$, age, APACHE III score, number of organs in failure according to the Brussels criteria, plateau pressure, plateau pressure missing (yes or no), and number of days in hospital prior to study entry. Our model includes an indicator of whether plateau pressure was missing or not. If that indicator is set to one then it is irrelevant what value we impute for the missing value of plateau pressure. Usually patients who are missing plateau pressure are on a different mode of ventilation, so the missing indicator method that we are using may be justified; in many cases, it would not be [48].

To compute the treatment effect adjusted for these covariates, a model is fit to the outcome of interest with treatment assignment and these seven covariates as the independent predictors using the entire cohort. From this model the probability of death at 60 days is computed for each patient from their observed covariates twice – once assuming they received treatment A, P_A, and once assuming they received treatment B, P_B. The difference between these, $\Delta = P_A - P_{B,}$ is the estimated treatment effect for the patient while the mean of Δ over the entire cohort is the estimated adjusted treatment effect. The variance of the estimated adjusted effect is the sum of two components [49,50]. The first component is the usual sample variance of Δ divided by the total N. The second is the mean of the model based estimates of the variance of Δ for each patient. The model-based variance is computed by the delta method [51] for each patient from the patient's covariates; the covariance matrix of the estimated model coefficients, and the model link function [52]. In every case we have done this, where the difference between this estimate of the treatment effect and the raw estimate has been small. In large randomized studies covariate imbalance does not have a large effect on the outcome of the trial.

Future perspective

It was interesting to compare the Clinical Coordinating Center grant application written in 1995 with the subsequent practice of the ARDS Network. Many of the proposals in our application were followed with great benefit. The most notable was the use of factorial trials. Looking forward to future trials in ARDS, a major issue is whether or not our current sample size is adequate. As ARDS mortality decreases it may become harder to find treatments that produce a substantial improvement. This is already true in pediatric trials and we have suggested methods of extending adult trials to children using Bayesian methods [53].

One approach is to shift to studies using VFD as a primary end point. This approach may allow us to show treatment effects at the cost of developing less compelling evidence that we have shown an unequivocal benefit of the new treatment. Long-term follow-up might also suggest new measures of treatment efficacy. An alternative is to have a network that can conduct larger trials. This will affect how we do trials; what is practical for 12 university centers with approximately 35 hospitals may not be practical for a network of 100 hospitals. Surely our governance would change. Currently most decisions are made by a committee of the whole with each institution getting one vote. We may also have to move to large simple trials with less emphasis on protocolized treatment and extensive data collection.

Furthermore, we may need a network that can move more quickly to take advantage of available opportunities and challenges. We responded to the H1N1 epidemic with two initiatives; one being a registry of H1N1 pediatric and adult cases treated in intensive care units and the second being a large simple trial of statins for H1N1. The registry was successful because we compensated for our late start by allowing the retrospective collection of cases at each institution. However, the clinical trial did not succeed because it was difficult to get timely funding.

There has been discussion in the network as to whether we could improve network operations using computerized decision support for some of the key aspects of our protocols such as mechanical ventilation and fluid management [54]. Currently we are considering a pilot project that would use these methods for half of our network and paper protocols for the other half. Our hope is that by using computerized decision support we could scale up the network

Executive summary

The definition of acute respiratory distress syndrome

- Acute lung injury is defined as a PaO_2/FiO_2 <300 mmHg, bilateral infiltrates and the requirement for positive pressure ventilation via an endotracheal tube not due to left atrial hypertension.
- Patients with PaO_2/FiO_2 between 200 and 300 mmHg had a mortality rate of 20.5% compared with 32.8% for those with values less than 200 mmHg.
- Using the wedge pressure to exclude heart failure is problematic as most patients do not have a pulmonary artery catheter and if one was in place many patients would be excluded.

Measurement of treatment efficacy in acute respiratory distress syndrome treatment

- The Acute Respiratory Distress Syndrome (ARDS) clinical trials network does not follow patients who go home with unassisted breathing. This probably has minimal effect on the 60-day mortality estimate although it may affect an estimate of 180-day mortality.
- Mortality should be compared between treatments using the proportion of people who died or a Kaplan–Meier estimate of this proportion and not the log-rank test.
- Ventilator-free days can be analyzed using a t-test. It can be used as a primary end point for ARDS clinical trials as long as we are convinced that we are not otherwise harming the patient.

Sample size for ARDS clinical trials

- The effect of a drop in the mortality rate of ARDS on sample size depends on whether you conceptualize differences as absolute or relative. If we wish to detect differences of 25%, say from 40 to 30% or from 20 to 15%, we will need larger sample sizes as we improve ARDS care.

Efficacy measurements based on lung function

- The change in lung physiology measures in Phase III clinical trials has been inconsistent so that these are not good efficacy measures.

Determination of the data to be collected

- Patients should be excluded from clinical trials when they have specific contraindications to the treatments. The ARDS Network does not collect reports of serious adverse events unless they were either unexpected or treatment related. We also do not collect extensive data on cointerventions or concomitant medications.

Stratification & randomization

- Stratification by hospital, which was not done by the ARDS Network, might increase power somewhat for trials that accrue to completion, but might reduce power for the early look at data.

Factorial trials & coenrollment

- Factorial designs and coenrollment, the testing of more than one treatment on the same patient, increased the number of questions that were addressed by the ARDS Network.

Early stopping rules

- Early stopping rules for efficacy and futility were an important part of ARDS Network trials.

Choice of a control group

- The ARDS Network used specific therapies as control groups rather than usual care. Usual care control groups are hard to interpret in ARDS clinical trials.

Strategy for cointerventions

- Cointerventions are treatments that are not being studied in the clinical trial. Controlling for these has a negligible effect on sample size. The ARDS Network only controlled those cointerventions where there was evidence that they improved patient outcomes.

Quality control of treatment administration

- The ARDS Network developed a method of controlling treatment administration by using checks of the patients' condition at a random time each day and produced reports as to whether the patient was on target at that time.

Controlling for prognostic factors

- Prognostic factors were controlled in a sensitivity analysis. A method was used that produced an estimate of the mortality difference as if all patients had been on both treatments.

Future perspective

- In the future we may have to conduct larger trials or change end points.
- The use of computer decision support and unified access to electronic records may make larger trials more feasible.
- Funding trials through a network has advantages over funding individual trials with research grants in that it allows for the rapid initiation of trials and encourages innovative trial designs, such as factorial designs and coenrollment.
- Current trials are conducted by the network, not by the industry. However, the industry has been generous in providing free drugs, supplies and services to the network.

without sacrificing treatment quality. The advent of some form of a unified medical record or unified access to hospital medical records would lead to considerable savings in the cost of data collection and also allow for larger trials.

One can ask whether clinical trial networks are a better way to fund clinical research than clinical trials funded by the research grant mechanism or funding by industry. The ARDS Network has been able to mount clinical trials much faster then trials funded by grants that often go through several peer review cycles before they begin. Furthermore, there has been a considerable cost savings through the development of a common infrastructure. Factorial trials and coenrolled trials are difficult to mount using a research grant mechanism due to the fact that the more questions one proposes, the more likely a reviewer will object to the study. The study of lisofylline [4] would not have been conducted by industry without the network. All the other studies were investigator initiated. The industry has been very cooperative with the ARDS Network. All but one of our drug studies had drugs and placebo supplied free to the network. In addition, the industry has provided services such as drug assays.

We hope this description of how the ARDS Network has handled statistical issues will be useful to others conducting clinical trials in acute diseases. Many of our decisions are a function of the disease we are treating and our capabilities as an organization. We hope that this description of our practice will be useful in other settings and to other clinical trialists in academia and industry who are conducting clinical trials in ARDS.

Acknowledgements

The authors would like to thank the invaluable technical assistance provided by Nora E Lavery.

Financial & competing interests disclosure

This work was supported by HHSN268200536179C. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:

■ of interest

■■ of considerable interest

1 Thompson B, Bernard G. ARDS Network (NHLBI) Studies: successes and challenges in ARDS clinical research. *Crit. Care. Clin.* 27(3), 459–468 (2011).

2 The Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N. Engl. J. Med.* 342(18), 1301–1308 (2000).

3 The Acute Respiratory Distress Syndrome Network. Ketoconazole for early treatment of acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. The ARDS Network. *JAMA* 283(15), 1995–2002 (2000).

4 NIH/NHLBI Acute Respiratory Distress Syndrome Network. Randomized, placebo-controlled trial of lisofylline for early treatment of acute lung injury and acute respiratory distress syndrome. *Crit. Care Med.* 30(1), 1–6 (2002).

5 Steinberg K, Hudson L, Goodman R *et al.* Efficacy and safety of corticosteroids for persistent acute respiratory distress syndrome. *N. Eng. J. Med.* 354(16), 1671–1684 (2006).

6 Brower R, Lanken P, MacIntyre N *et al.* Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N. Engl. J. Med.* 351(4), 327–336 (2004).

7 Wheeler AP, Bernard GR, Thompson BT *et al.* Pulmonary-artery versus central venous catheter to guide treatment of acute lung injury. *N. Engl. J. Med.* 354(21), 2213–2224 (2006).

8 Wiedemann HP, Wheeler AP, Bernard GR *et al.* Comparison of two fluid-management strategies in acute lung injury. *N. Engl. J. Med.* 354(24), 2564–2575 (2006).

9 National Heart Lung and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network, Matthay M, Brower R *et al.* Randomized, placebo-controlled clinical trial of an aerosolized beta-2 agonist for treatment of Acute Lung Injury. *Am. J. Respir. Crit. Care Med.* 184(5), 561–568 (2011).

10 Rice T, Wheeler A, Thompson B *et al.* Enteral omega-3 fatty acid, gamma-linolenic acid, and antioxidant supplementation in acute lung injury. *JAMA* 306(14), 1574–1581 (2011).

11 Murray JF, Matthay MA, Luce JM, Flick MR. An expanded definition of the adult respiratory distress syndrome. *Am. Rev. Respir. Dis.* 138(3), 720–723 (1988).

12 Bernard GR, Luce JM, Sprung CL *et al.* High-dose corticosteroids in patients with the adult respiratory distress syndrome. *N. Engl. J. Med.* 317(25), 1565–1570 (1987).

13 Steinberg KP, Mitchell DR, Maunder RJ *et al.* Safety of bronchoalveolar lavage in patients with adult respiratory distress syndrome. *Am. Rev. Respir. Dis.* 148(3), 556–561 (1993).

14 Bernard GR, Artigas A, Brigham KL *et al.* The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. *Am. J. Respir. Crit. Care Med.* 149(3 Pt 1), 818–824 (1994).

15 Connors AJ, Speroff T, Dawson N *et al.* The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *JAMA* 276(11), 889–897 (1996).

16 Clermont G, Kong L, Weissfeld L *et al.* The effect of pulmonary artery catheter use on costs and long-term outcomes of acute lung injury. *PLoS ONE* (2011) (In Press).

17 Kalbfleisch J, Prentice R. *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York, NY, USA (1980).

18 Schoenfeld D. Survival methods, including those using competing risk analysis, are not appropriate for intensive care unit outcome studies. *Crit. Care* 10(1), 103 (2006).

19 Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 68, 316–319 (1981).

20 Zelen M. The analysis of several 2X2 contingency tables. *Biometrika* 58(1), 127–137 (1971).

21 Betensky RA, Schoenfeld DA. Nonparametric estimation in a cure model with random cure times. *Biometrics* 57(1), 282–286 (2001).

22 Schoenfeld D, Bernard G, ARDS Network. Statistical evaluation of ventilator-free days as an efficacy measure in clinical trials of treatments for acute respiratory distress syndrome. *Crit. Care Med.* 30(8), 1772–1777

(2002).

- **Introduces ventilator-free days as a primary end point for acute respiratory distress syndrome trials and discusses sample size considerations and the interpretation of the results.**

23 Miller R. *Beyond ANOVA, Basics of Applied Statistics*. John Wiley & Sons, New York, NY, USA (1986).

24 Peto R, Pike MC, Armitage P *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34(6), 585–612 (1976).

25 Knaus WA, Wagner DP, Draper EA *et al.* The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 100(6), 1619–1636 (1991).

26 Bull T, Clark B, McFann K, Moss M. Pulmonary vascular dysfunction is associated with poor outcomes in patients with acute lung injury. *Am. J. Respir. Crit. Care Med.* 182(9), 1123–1128 (2010).

27 Grams M, Estrella M, Coresh J *et al.* Fluid balance, diuretic use, and mortality in acute kidney injury. *Clin. J. Am. Soc. Nephrol.* 6(5), 966–973 (2011).

28 Rice T, Ware L, Haponik E *et al.* Vascular pedicle width in acute lung injury: correlation with intravascular pressures and ability to discriminate fluid status. *Crit. Care* 15(2), R86 (2011).

29 Wang W, Scharfstein D, Wang C *et al.* Estimating the causal effect of low tidal volume ventilation on survival in patients with acute lung injury. *J. R. Stat. Soc.* 60(4), 475–496 (2011).

30 Bohn MJ, Babor TF, Kranzler HR. The Alcohol Use Disorders Identification Test (AUDIT): validation of a screening instrument for use in medical settings. *J. Stud. Alcohol* 56(4), 423–432 (1995).

31 Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 2(7872), 81–84 (1974).

32 Bernard G. The Brussels Score. *Sepsis* 1(1), 43–44 (1997).

33 Cook R, Farewell V. Multiplicity considerations in the design and analysis of clinical trials. *J. R. Statist. Soc. A* 159, 93–110 (1996).

34 Byar DP, Schoenfeld DA, Green SB *et al.* Design considerations for AIDS trials. *N. Engl. J. Med.* 323(19), 1343–1348 (1990).

- ■ **Takes a strategic view of clinical trials in a particular disease area. Many of the ideas in this paper are crucial to the development of clinical trial strategies in other disease, such as acute respiratory distress syndrome, particularly the discussion of factorial design and the introduction of the idea of coenrollment.**

35 Demets D, Lan G. An overview of sequential methods and their application in clinical trials. *Commun. Stat. Theory Methods* 13(19), 2315–2338 (1984).

36 O’Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 35(3), 549–556 (1979).

37 Schoenfeld D. A simple algorithm for designing group sequential clinical trials. *Biometrics* 57(3), 972–974 (2001).

- **Describes the futility boundaries that were used in several Acute Respiratory Distress Syndrome Network trials and focuses.**

38 Jennison C, Turnbull B. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, New York, NY, USA (2000).

- **Best reference on group sequential trials.**

39 DeMets D, Ware L. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69(3), 661–663 (1982).

40 Thompson BT, Hayden D, Matthay MA, Brower R, Parsons PE. Clinicians’ approaches to mechanical ventilation in acute lung injury and ARDS. *Chest* 120(5), 1622–1627 (2001).

41 Hickling KG, Henderson SJ, Jackson R. Low mortality associated with low volume pressure limited ventilation with permissive hypercapnia in severe adult respiratory distress syndrome. *Intensive Care Med.* 16(6), 372–377 (1990).

42 Hickling KG, Walsh J, Henderson S, Jackson R. Low mortality rate in adult respiratory distress syndrome using low-volume, pressure-limited ventilation with permissive hypercapnia: a prospective study. *Crit. Care Med.* 22(10), 1568–1578 (1994).

43 Thompson BT, Schoenfeld D. Usual care as the control group in clinical trials of nonpharmacologic interventions. *Proc. Am. Thorac. Soc.* 4(7), 577–582 (2007).

44 Eichacker P, Gerstenberger E, Banks S, Cui X, Natanson C. Meta-analysis of acute lung injury and acute respiratory distress syndrome trials testing low tidal volumes. *Am. J. Respir. Crit. Care Med.* 166(11), 1510–1514 (2002).

45 Spragg R, Bernard G, Checkley W *et al.* Beyond mortality: future clinical research in acute lung injury. *Am. J. Respir. Crit. Care Med.* 181(10), 1121–1127 (2010).

46 Beach M, Meier P. Choosing covariates in the analysis of clinical trials. *Control. Clin. Trials* 10(4), 161–175 (1989).

47 Senn S. Controversies concerning randomization and additivity in clinical trials. *Stat. Med.* 23(24), 3729–3753 (2004).

48 Jones M. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J. Am. Stat. Ass.* 91(443), 222–230 (1996).

49 Chang IM, Gelman R, Pagano M. Corrected group prognostic curves and summary statistics. *J. Chronic. Dis.* 35(8), 669–674 (1982).

50 Graubard BI, Korn EL. Predictive margins with survey data. *Biometrics* 55(2), 652–659 (1999).

51 Schervish M. *Theory of Statistics*. Springer-Verlag. NY, USA (1995).

52 McCullagh P, Nelder J. *Generalized linear models* (*2nd Edition*). Chapman & Hall, New York, NY, USA (1989).

53 Schoenfeld D, Zheng H, Finkelstein D. Bayesian design using adult data to augment pediatric trials. *Clinical Trials* 6(4), 297–304 (2009).

54 Morris A, Sorenson D, Warner H *et al.* Bedside computerized study protocols for an ARDS Network clinical trial. *Am. J. Resp. Crit. Care Med.* 171(Abstr.) (2005).

■ Websites

101 Statistical considerations for clinical trials and scientific experiments. www.hedwig.mgh.harvard.edu/sample_size/size.html

102 Reporting serious problems to the FDA. What is a serious adverse event? www.fda.gov/safety/medwatch/howtoreport/ucm053087.htm